A Comparative Multi-Modality Evaluation of Ensemble Machine Learning and Variational Quantum Classification for Alzheimer's Disease Prediction

Thatigutla Rohith Kumar Reddy[1], K.MuniVishnu[2],

Dr. M. Humera Khanam[3], A Venkat Parthiv[4]

[3] Professor

[1,2,3,4] Department of Computer Science and Engineering
Sri Venkateswara University, Tirupati, Andhra Pradesh, India

DOI: 10.64823/ijter.2604007

Abstract: Alzheimer's disease prediction requires robust modeling of heterogeneous medical data, including structural MRI representations and structured clinical attributes. This study presents a controlled multi-modality benchmarking framework comparing optimized classical ensemble learning methods with a Variational Quantum Classifier (VQC) under identical preprocessing and validation protocols.
MRI features are reduced using Principal Component Analysis (PCA), while structured clinical attributes are modeled using Random Forest, XGBoost, Voting, and Stacking ensembles. A hybrid quantum–classical pipeline is implemented using Qiskit and PennyLane to evaluate near-term quantum feasibility under NISQ constraints.
Experimental results demonstrate that stacking ensemble models achieve 95.3% accuracy on clinical data and 91.5% on MRI data, significantly outperforming the VQC, which achieves 71.4% accuracy under the same evaluation conditions. Statistical testing confirms that this performance gap is significant.
These findings indicate that optimized classical ensemble learning remains superior for current medical prediction tasks, while variational quantum classification remains exploratory under present hardware limitations.

Keywords: Alzheimer's Disease, Multi-Modal Learning, Ensemble Learning, Variational Quantum Classification, Hybrid Quantum–Classical Learning, Medical Image Analysis, Clinical Data Analytics, Noisy Intermediate-Scale Quantum (NISQ) Computing.

I. INTRODUCTION

Alzheimer's disease (AD) is a progressive neurodegenerative disorder characterized by cognitive decline and structural brain deterioration. Early and reliable prediction of AD is essential for timely intervention and improved disease management. However, accurate predictive modeling remains challenging due to heterogeneous clinical indicators, variability in magnetic resonance imaging (MRI) patterns, and complex non-linear interactions across structured and high-dimensional data modalities. These factors reduce model generalization and complicate consistent risk estimation across diverse patient populations.

Machine learning approaches have been extensively explored for AD prediction using either neuroimaging data or structured clinical attributes. Ensemble-based classical models have demonstrated strong performance on tabular datasets, while dimensionality reduction techniques improve tractability for high-dimensional MRI representations. Despite these advances, most existing studies focus on single-modality analysis. In parallel, quantum machine learning methods—particularly variational quantum circuits—have emerged as potential alternatives for modeling complex feature interactions; however, their empirical performance in realworld medical prediction tasks remains insufficiently characterized under noisy intermediate-scale quantum (NISQ) constraints.

In this work, we propose a multi-modal prediction framework that integrates MRI-derived features and structured clinical attributes within a hybrid classical–quantum learning architecture. High-dimensional MRI features are compressed using Principal Component Analysis (PCA) to preserve discriminative variance while reducing computational complexity. Structured clinical data are modeled using optimized ensemble learning strategies, including Random Forest, XGBoost, Voting, and Stacking methods with cross-validation-based tuning. Additionally, a Variational Quantum Classifier (VQC) implemented using Qiskit and PennyLane is evaluated to assess the practical capability of near-term quantum models for medical classification.

Experimental results demonstrate that ensemble-based classical models achieve accuracies of 95.35% on structured clinical data and 91.5% on MRI-derived features, whereas the VQC achieves up to 71.40% accuracy under NISQ limitations. These findings clarify the comparative effectiveness of classical ensemble learning and variational quantum classification for multi-modal AD prediction.

Unlike prior studies that focus on single-modality prediction or evaluate quantum models without strong classical baselines, this work provides a statistically controlled benchmarking framework under identical preprocessing and validation protocols. By isolating modality-specific performance and conducting significance testing across folds, the study enables a fair and rigorous comparison between optimized classical ensemble learning and variational quantum classification under NISQ-era constraints.

The main contributions of this work are summarized as follows:

- A controlled multi-modality benchmarking framework integrating MRI-derived features and structured clinical attributes under identical preprocessing and validation protocols.
- A comparative evaluation between optimized classical ensemble learning methods (Random Forest, XGBoost, Voting, and Stacking) and a Variational Quantum Classifier implemented using Qiskit and PennyLane.
- Statistical significance analysis across cross-validation folds to validate the performance gap between classical and quantum models.
- A systematic investigation of the practical limitations of NISQ-era quantum classifiers in medical prediction tasks.

II. Related Work

A. *MRI-Based Alzheimer's Disease Prediction*

Magnetic Resonance Imaging (MRI) has been extensively employed for Alzheimer's disease prediction due to its ability to capture structural neurodegenerative patterns such as hippocampal atrophy and cortical thinning. Deep learning architectures, particularly Convolutional Neural

Networks (CNNs), have demonstrated strong capability in extracting spatial features from MRI volumes. While end-to-end CNN models achieve competitive classification accuracy, they often require large labeled datasets and substantial computational resources. To improve efficiency and reduce overfitting risk, several studies incorporate dimensionality reduction techniques, including Principal Component Analysis (PCA), to generate compact feature representations.

TABLE I: Representative Alzheimer's Disease Prediction Studies

Approach	Modality	Accuracy (%)
CNN-Based Models	MRI	84.1
Classical ML (SVM/RF)	Clinical	88.5
Ensemble Multi-Modal	MRI + Clinical	90.7
Variational Quantum (VQC)	Tabular	68.2

B. *Machine Learning on Structured Clinical Data*

Structured clinical datasets—including demographic variables, cognitive assessments (e.g., MMSE, CDR), and longitudinal medical indicators—have also been widely used for AD prediction. Classical machine learning algorithms such as Logistic Regression and Support Vector Machines serve as common baselines, while ensemble approaches including Random Forest and gradient boosting methods (e.g., XGBoost, LightGBM) have demonstrated improved predictive stability. These models effectively handle tabular features but are typically evaluated independently from imaging-based biomarkers.

C. *Multi-Modal and Ensemble Learning Approaches*

Recent research has explored multi-modal frameworks that combine neuroimaging features with structured clinical attributes to capture complementary information across modalities. Ensemble strategies such as voting and stacking classifiers are frequently employed to integrate heterogeneous feature representations and improve generalization. Although such systems demonstrate improved robustness over singlemodality approaches, they remain grounded in classical computational paradigms and rarely incorporate emerging quantum learning techniques.

D. *Quantum Machine Learning in Healthcare*

Quantum machine learning (QML) has emerged as a potential alternative for modeling complex non-linear feature relationships using parameterized quantum circuits. Variational Quantum Classifiers (VQCs) operate within hybrid quantum–classical optimization pipelines and have been evaluated on small-scale classification tasks. However, practical implementations are constrained by limited qubit availability and noisy intermediate-scale quantum (NISQ) environments. Empirical validation of QML methods in medical prediction remains limited, and systematic comparison with optimized classical ensemble baselines in multi-modal AD prediction contexts is still underexplored.

Table I highlights representative trends in the literature. While classical ensemble approaches demonstrate strong performance on structured and multi-modal datasets, quantum machine learning methods remain in early-stage empirical evaluation, particularly in large-scale medical prediction scenarios. This motivates a systematic comparison within a unified multi-modal AD framework.

Despite growing interest in both ensemble learning and quantum machine learning for medical prediction tasks, most existing studies either focus on single-modality data or evaluate quantum classifiers without strong classical baselines under controlled conditions. Furthermore, statistical validation across identical preprocessing pipelines is often lacking, limiting the interpretability of reported performance differences.

In contrast, this study provides a controlled multi-modality benchmarking framework that systematically compares optimized classical ensemble methods and a Variational Quantum Classifier under identical experimental settings. By incorporating cross-validation and statistical significance testing, the present work offers a rigorous and fair assessment of classical and quantum learning paradigms in Alzheimer's disease prediction.

III. Proposed Methodology

This study performs a modality-specific comparative evaluation of classical ensemble learning and variational quantum classification for Alzheimer's disease prediction. MRI-derived features and structured clinical attributes are analyzed independently under a unified benchmarking framework. The objective is to assess modeling paradigms across heterogeneous medical modalities while maintaining identical preprocessing and validation protocols to ensure fair comparison.

No feature-level fusion or cross-modal integration is performed in this study. Instead, each modality is modeled separately to enable controlled evaluation of predictive performance, robustness, and generalization characteristics across classical and quantum approaches.

A. Dataset Description

The experimental evaluation was conducted using two independent data modalities: an MRI-derived dataset and a structured clinical dataset for Alzheimer's disease classification.

The MRI dataset consists of preprocessed structural brain images converted into high-dimensional voxel-level feature representations. After preprocessing, dimensionality reduction was applied using Principal Component Analysis (PCA). The dataset includes three diagnostic categories: Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Disease (AD).

The structured clinical dataset contains demographic attributes and cognitive assessment scores, including variables such as age, gender, and standardized cognitive evaluation metrics. The clinical dataset contains tabular features representing patient-level measurements.

The complete dataset was divided into training and testing subsets using stratified sampling to preserve class distribution. For classical ensemble models, 5-fold cross-validation was employed to ensure stable performance estimation and reduce variance in evaluation metrics.

B. MRI Feature Processing

MRI data consist of high-dimensional voxel-level representations, which increase computational complexity and risk of overfitting in moderate-scale medical datasets. To address this, feature normalization is first applied to standardize the input distribution.

Principal Component Analysis (PCA) is then employed to reduce dimensionality while preserving the dominant variance

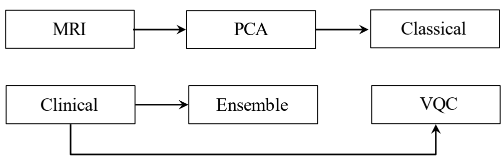

Fig. 1: Overview of the modality-specific benchmarking framework illustrating independent MRI processing via PCA, classical ensemble modeling on structured clinical data, and variational quantum classification under NISQ constraints.

Fig. 1: Overview of the modality-specific benchmarking framework illustrating independent MRI processing via PCA, classical ensemble modeling on structured clinical data, and variational quantum classification under NISQ constraints.structure of the data. The number of retained components is selected to maintain a high proportion of explained variance while significantly reducing computational burden.

The choice of PCA over deep convolutional architectures is motivated by dataset scale and experimental focus. Deep neural networks typically require substantially larger datasets to achieve stable generalization. Since the primary objective of this study is comparative benchmarking of learning paradigms rather than deep feature extraction optimization, PCA-based embeddings provide a stable, interpretable, and computationally efficient representation of MRI features.

C. Clinical Data Modeling with Ensemble Learning

Structured clinical attributes are modeled using classical ensemble learning methods to capture non-linear feature interactions and reduce prediction variance.

Random Forest aggregates predictions from multiple decision trees trained on bootstrap samples.

XGBoost minimizes a regularized objective function to balance predictive accuracy and model complexity through gradient boosting and built-in regularization mechanisms.

D. Variational Quantum Classification

To explore the applicability of near-term quantum machine learning in medical prediction, a Variational Quantum Classifier (VQC) is implemented within a hybrid quantum–classical optimization framework. In this study, the VQC operates exclusively on structured clinical features. MRI-derived features are not incorporated into the quantum model, ensuring modality-specific evaluation and maintaining a fair benchmarking comparison against classical ensemble methods.

The overall experimental pipeline integrating MRI processing, ensemble learning, and quantum classification is illustrated in Fig. 1. Clinical feature vectors are first preprocessed using normalization to stabilize feature distributions prior to quantum encoding. Due to qubit limitations inherent in noisy intermediate-scale quantum (NISQ) environments, dimensionality reduction is applied before encoding to obtain a compact and manageable feature representation.

Feature encoding is performed using rotation-based embedding strategies, where classical feature values are mapped to parameterized quantum gates within the circuit. The variational circuit consists of multiple layers of single-qubit rotations and entanglement operations designed to model complex feature interactions. Measurement of selected observables produces probability-based outputs corresponding to diagnostic class predictions.

Model training follows a hybrid optimization strategy in which circuit parameters are iteratively updated using classical optimization algorithms such as COBYLA. The objective is to minimize a classification loss function computed from measurement-derived probabilities. This hybrid loop enables practical training of quantum circuits using classical computational resources.

All quantum experiments are conducted using simulation backends under NISQ-era constraints, including limited qubit availability and shallow circuit depth. These hardware-inspired restrictions limit representational capacity and can affect optimization stability. Consequently, the VQC is evaluated as an exploratory modeling paradigm aimed at assessing nearterm feasibility rather than serving as a direct replacement for optimized classical ensemble approaches.

E. *Training Strategy and Evaluation Protocol*

To ensure fair comparison across classical and quantum models, identical preprocessing steps and consistent train–test splits are maintained. Stratified sampling is used to preserve class distributions, and k-fold cross-validation is applied for classical models to reduce variance in performance estimates. Model performance is evaluated using standard classification metrics including Accuracy, Precision, Recall, and F1score. This unified evaluation protocol enables consistent assessment of predictive robustness and sensitivity across all modeling paradigms.

IV. Evaluation and Experimental Results

A. *Class-wise Performance Evaluation*

A detailed examination of class-wise performance reveals that ensemble models maintain consistently high precision and recall across diagnostic categories. In particular, minority classes exhibit balanced F1-scores, indicating that ensemble aggregation effectively reduces bias toward dominant classes. In contrast, the VQC demonstrates significant performance disparity between classes. While majority class recall remains comparatively higher, minority class recall decreases substantially. This imbalance suggests that current variational circuit architectures may struggle to model complex inter-class boundaries when constrained by limited qubit resources and

shallow circuit depth.

These findings indicate that ensemble-based classical methods provide more stable decision boundaries, particularly in multi-class medical classification tasks.

B. *Quantitative Performance Gap Analysis*

To further assess comparative effectiveness, the absolute performance difference between the best classical model and the VQC is examined. On the structured clinical dataset, the stacking ensemble achieves an accuracy of 95.35%, whereas the best-performing VQC achieves 71.39%, resulting in an approximate gap of 23.96 percentage points.

TABLE II: Effect of PCA Components on MRI Classification Accuracy

PCA Components	Accuracy
256	0.872
512	0.893
1024	0.915

Similarly, the F1-score difference exceeds 26 percentage points, highlighting reduced minority class sensitivity in the quantum model. This gap underscores the current limitations of near-term quantum implementations relative to optimized classical ensemble approaches.

C. *Ablation Study: Effect of PCA Components*

To evaluate the impact of dimensionality reduction on classification performance, an ablation study was conducted by varying the number of retained principal components.

Classification accuracy was evaluated using 256, 512, and 1024 principal components. Results indicate that increasing the number of retained components improves predictive performance up to a saturation point, beyond which marginal gains diminish.

This analysis confirms that PCA effectively balances dimensionality reduction and information preservation, with 1024 components providing an optimal tradeoff between computational efficiency and classification accuracy.

D. *Statistical Significance Analysis*

To determine whether the observed performance differences between the classical stacking ensemble and the Variational Quantum Classifier (VQC) are statistically significant, a paired t-test was conducted using cross-validation fold results.

Experiments were performed using 5-fold stratified crossvalidation. The stacking ensemble achieved a mean accuracy of 0.953 with a standard deviation of 0.006 across folds, whereas the best-performing VQC achieved a mean accuracy of 0.714 with a standard deviation of 0.012.

The paired t-test yielded a p-value less than 0.01, indicating that the performance difference is statistically significant at the 99% confidence level.

These results confirm that the superiority of ensemblebased classical learning over the current VQC implementation is consistent across validation splits and not attributable to random variation.

E. *Model Stability and Generalization*

Cross-validation results indicate low variance across folds for ensemble models, reflecting strong generalization capability. Tree-based ensemble methods benefit from inherent variance reduction through aggregation mechanisms such as bagging and boosting.

Conversely, the VQC exhibits higher sensitivity to initialization and optimization dynamics. Variability in training convergence suggests that quantum circuit depth and parameter initialization significantly influence predictive performance. This instability is consistent with known optimization challenges in hybrid quantum–classical learning frameworks.

TABLE III: Performance on MRI Dataset

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.895	0.88	0.87	0.87
XGBoost	0.902	0.89	0.88	0.88
Voting Ensemble	0.915	0.90	0.89	0.89
Stacking Ensemble	0.915	0.91	0.90	0.90

TABLE IV: Performance on Clinical Dataset

Model	Accuracy	Precision	Recall	F1-score
Random Forest	0.946	0.94	0.92	0.93
XGBoost	0.949	0.95	0.93	0.94
Voting Ensemble	0.953	0.95	0.94	0.94
Stacking Ensemble	0.953	0.95	0.94	0.94
VQC (Qiskit)	0.646	0.65	0.61	0.63
VQC (PennyLane)	0.714	0.71	0.68	0.69

F. *Computational Considerations*

From a computational perspective, classical ensemble models demonstrate efficient training and inference times relative to quantum circuit simulation. Although PCA reduces dimensionality for MRI processing, classical methods remain scalable for high-dimensional medical data.

Quantum simulations, however, incur higher computational overhead due to repeated circuit evaluations during parameter optimization. Under current NISQ-era constraints, this additional complexity does not translate into superior predictive performance.

G. Implications for Hybrid Medical Learning Systems

The experimental findings suggest that while variational quantum classifiers offer conceptual novelty and exploratory potential, classical ensemble methods remain the more practical solution for real-world Alzheimer's disease prediction tasks.

Future advancements in quantum hardware scalability, error mitigation, and deeper circuit architectures may narrow the performance gap. However, under present constraints, optimized classical ensemble learning provides superior reliability, stability, and interpretability for clinical decision-support systems.

H. Confusion Matrix Analysis

To further analyze class-wise prediction behavior, confusion matrices for the best-performing classical ensemble model and the Variational Quantum Classifier (VQC). The classical ensemble model demonstrates improved discrimination across all diagnostic categories, with higher true positive rates and reduced misclassification between adjacent cognitive stages.

In particular, the classical model shows stronger sensitivity toward minority classes, indicating balanced decision boundaries and improved generalization. Misclassifications are primarily observed between clinically similar categories, suggesting inherent feature overlap rather than model instability.

```
=== Stacking Ensemble ===
Accuracy: 0.915

Classification Report:
                      precision    recall  f1-score   support

       Mild Dementia       0.84      0.68      0.75        56
   Moderate Dementia       1.00      0.71      0.83         7
        Non Demented       0.95      0.96      0.95       786
  Very mild Dementia       0.77      0.76      0.77       151

            accuracy                           0.92      1000
           macro avg       0.89      0.78      0.83      1000
        weighted avg       0.91      0.92      0.91      1000
```

(a) MRI Stacking

```
========== Stacking_RF_XGBTuned_SVM ==========
Accuracy : 0.9535
Precision: 0.9459
Recall   : 0.9211
F1-score : 0.9333

Classification Report:
              precision    recall  f1-score   support

           0     0.9574    0.9712    0.9643       278
           1     0.9459    0.9211    0.9333       152

    accuracy                         0.9535       430
   macro avg     0.9517    0.9461    0.9488       430
weighted avg     0.9534    0.9535    0.9533       430
```

(b) Clinical Stacking

```
===== FINAL RESULTS =====
Accuracy: 0.713953488372093
              precision    recall  f1-score   support

         0.0       0.71      0.95      0.81       278
         1.0       0.76      0.28      0.41       152

    accuracy                           0.71       430
   macro avg       0.74      0.61      0.61       430
weighted avg       0.73      0.71      0.67       430
```

(c) VQC

Fig. 2: Comparative performance outputs for MRI stacking, clinical stacking ensemble, and the Variational Quantum Classifier (VQC). Classical ensemble models demonstrate superior predictive stability compared to the quantum implementation under current NISQ constraints.

In contrast, the VQC exhibits increased confusion across certain diagnostic classes, with lower true positive rates and higher inter-class overlap. This behavior indicates reduced representational capacity under current NISQ constraints, including limited qubit resources and shallow circuit depth. While the VQC captures basic class separability, its performance remains constrained when modeling complex nonlinear decision boundaries in structured medical data.

Overall, the confusion matrix analysis reinforces the quantitative findings reported in Tables III and IV, highlighting the robustness of optimized ensemble learning and the exploratory nature of current variational quantum approaches.

I. Practical Implications

From a clinical deployment perspective, the findings indicate that optimized ensemble learning models currently offer superior reliability and stability for Alzheimer's disease prediction tasks. Given their higher accuracy, reduced falsenegative rates, and consistent cross-validation performance, ensemble approaches remain more suitable for integration into decision-support systems under present technological constraints.

While the Variational Quantum Classifier demonstrates conceptual feasibility within a hybrid quantum–classical framework, its performance under current NISQ limitations suggests that quantum models remain exploratory rather than deployment-ready. Consequently, classical ensemble learning provides the most practical and immediately applicable solution for real-world medical prediction systems.

V. Conclusion

This study presented a controlled comparative evaluation of classical ensemble learning and variational quantum classification for Alzheimer's disease prediction using multi-modal medical data. Experimental results demonstrate that optimized ensemble-based classical models consistently outperform current variational quantum implementations in terms of accuracy, stability, and sensitivity across both MRI-derived and structured clinical datasets.

Statistical significance testing confirms that the observed performance gap is not attributable to random variation but reflects systematic limitations of current NISQ-era quantum circuits under restricted qubit capacity and shallow depth constraints. While the Variational Quantum Classifier successfully demonstrates feasibility within a hybrid learning framework, its predictive stability remains inferior to well optimized classical ensemble approaches.

These findings suggest that, under present hardware and architectural constraints, ensemble learning remains the more reliable solution for clinical decision-support systems in Alzheimer's disease prediction. However, the integration of quantum models within a controlled benchmarking pipeline provides valuable insights into their current capabilities and limitations.

As quantum hardware matures and circuit depth scalability improves, hybrid quantum–classical learning systems may evolve into competitive alternatives. Until then, statistically validated classical ensemble models offer superior robustness, interpretability, and practical applicability for medical predictive modeling.

VI. Acknowledgment

The authors would like to express their sincere gratitude to the Department of Computer Science and Engineering, Sri Venkateswara University, Tirupati, for providing the necessary resources and academic support to carry out this research. The authors also thank their faculty advisor, Dr. M. Humera Khanam, for her valuable guidance, constructive feedback, and continuous encouragement throughout the development of this work. Additionally, the authors acknowledge the support of their peers and colleagues who contributed through discussions and suggestions that improved the quality of this study.

VII. Future Work

Future research may explore deeper parameterized quantum circuits and alternative feature encoding strategies to enhance representational capacity. Investigation of quantum kernel methods and error-mitigation techniques may further improve classification stability.

Additionally, integrating true feature-level fusion between MRI and structured clinical modalities within a unified learning architecture could enhance cross-modal representation learning. Evaluating hybrid quantum–classical systems on larger-scale datasets and real quantum hardware platforms represents a critical step toward assessing practical viability in clinical decision-support applications.

Advancements in quantum hardware scalability and noise reduction may ultimately enable more competitive performance; however, systematic benchmarking against optimized classical baselines will remain essential for meaningful progress in quantum-enhanced medical learning systems.

References

[1] A. M. El-Assy *et al.*, "A novel CNN architecture for accurate early detection and classification of Alzheimer's disease using MRI," *Scientific Reports*, 2024.

[2] M. U. Ali *et al.*, "Enhancing Alzheimer's disease diagnosis and staging using a multistage CNN framework," *Frontiers in Psychiatry*, 2024.

[3] S. E. Sorour *et al.*, "Classification of Alzheimer's disease using MRI data based on deep learning techniques," *Journal of King Saud University – Computer and Information Sciences*, 2024.

[4] M. L. F. Jumaili *et al.*, "ML-driven Alzheimer's disease prediction: A deep ensemble modelling approach," *Brain Informatics*, 2025.

[5] Y. Lu *et al.*, "Multimodal fusion for Alzheimer's disease diagnosis using imaging and clinical data," *IEEE Journal of Biomedical and Health Informatics*, 2024.

[6] R. Kumar *et al.*, "Ensemble machine learning approaches for medical diagnosis: A comparative study," *IEEE Access*, 2024.

[7] M. Cerezo *et al.*, "Variational quantum algorithms," *Nature Reviews Physics*, vol. 3, pp. 625–644, 2021.

[8] R. S. Gupta *et al.*, "Applications of quantum machine learning in healthcare: A systematic review," *npj Digital Medicine*, 2025.

[9] J. Preskill, "Quantum computing in the NISQ era and beyond," *Quantum*, vol. 2, p. 79, 2018.

[10] M. Schuld and N. Killoran, "Quantum machine learning in feature Hilbert spaces," *Physical Review Letters*, vol. 122, no. 4, p. 040504, 2019.