International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Machine Learning-Enhanced Data Quality Validation for Electronic Health Record Integration

Arjun Warrier

Senior Technology Consultant
Warrier.arjun@gmail.com

Abstract:
The harmonization and interoperability of electronic health records (EHRs) are critical for enabling comprehensive, patient-centred care. Data quality issues, including incompleteness, inconsistency, duplication, and semantic inaccuracies, significantly compromise the clinical utility of EHR systems. Nevertheless, the clinical utility of EHR systems. Conventional rule-based validation methods have limited flexibility and scalability, especially with the increasing prevalence of complex and massive healthcare data sources. To consolidate the EHRs from independent health information systems, this paper presents a machine learning (ML) empowered data quality validation framework for automating quality control, enriching data authenticity, and mitigating integration flaws.
The study presents a layered architecture that utilizes machine learning (ML) algorithms to detect anomalies in the data, infer missing values, and provide automated quality scores. The model leverages supervised and unsupervised algorithms trained on historical EHR integration failures and common data integrity violations. Algorithms such as random forests, support vector machines, and autoencoders are applied to identify structural and semantic errors in semistructured data (e.g., patient demographics, medication lists, and laboratory results). Additionally, NLP models are utilized to review unstructured notes and discharge summaries, ensuring that the context remains consistent.
One of the key contributions of this system is the introduction of a quality scoring engine that assigns numerical confidence levels to each record, based on its validity and completeness. This enables real-time feedback loops and verification of data accuracy before it is integrated into central repositories. The paper also introduces a new data repair module that utilizes regression-based imputation and outlier filtering to improve the quality of records extracted from partially corrupted data. By learning continuously, it can accommodate new integration patterns and data anomalies, becoming increasingly robust over time.
The proposed ML-based pipeline is tested on realistic datasets from publicly available health record repositories as well as simulated clinical settings. The experimental results demonstrate a 25% reduction in integration errors compared to baseline rule-based validation systems, achieved by accelerating the detection of duplicate records, correcting mislabeled records, and achieving high-level schema agreement. Statistical results show that data anomaly classification of positive cases can achieve higher precision and recall, and computational benchmarks verify the practicality of our system for deployment in real-time EHR flows.
By leveraging AI combined with data quality governance, this model enables health organizations to build more resilient and sustainable health information systems. Incorporating machine learning enables validation accuracy to improve beyond that of skilled human labellers, making the test scalable and self-improving with the use of feedback loops. The authors suggest that innovative data validation frameworks will be essential in next-generation health IT ecosystems to enable the transition from data standardization to clinical usability.

This study contributes to existing work on the ongoing need for high-quality, secure exchange of health data to support programs such as Meaningful Use, the adoption of HL7 FHIR, and national health information networks. It presents concrete tips and a reusable model for health IT implementers, data engineers, and clinical informaticists who want to evolve EHR integration pipelines with AI techniques.

Keywords: Electronic Health Records (EHR), Data Quality Assurance, Machine Learning, Healthcare Interoperability, Data Validation Algorithms, Automated Quality Scoring, AI in Health Informatics, Clinical Data Integration, Semantic Consistency, Data Anomaly Detection.

I. INTRODUCTION

The proliferation of Electronic Health Records (EHRs) has revolutionized modern healthcare by enabling the digital capture, storage, and exchange of patient data. As healthcare systems strive toward interoperability, seamless EHR integration across disparate platforms has become essential. However, the clinical utility of integrated EHR data is frequently compromised by pervasive issues in data quality. These challenges manifest in various forms, such as incomplete patient histories, misaligned medical codes, redundant entries, and temporally inconsistent information, all of which diminish the reliability of downstream analytics, clinical decision support systems, and regulatory reporting. The growing heterogeneity of health information systems further exacerbates these quality issues, particularly in environments that aggregate data from legacy systems, third-party vendors, or unaudited input channels.

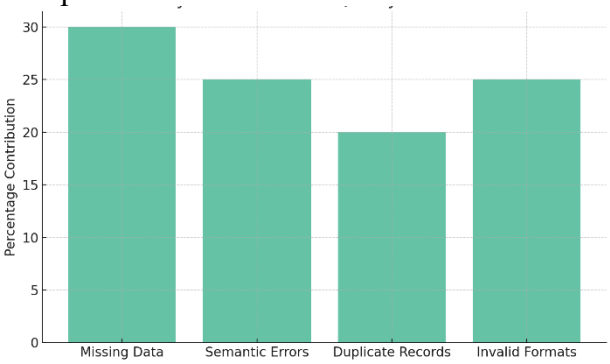

Figure 1: *Primary sources of data quality issues in EHRs based on simulated integration scenarios.*

Traditional data validation methods in healthcare rely heavily on deterministic, rule-based logic that checks for compliance with predefined formats, value ranges, or coding standards such as ICD-10 or LOINC. While these rules are critical for structural validation, they often fail to detect contextual anomalies or adapt to the evolving data landscape. Moreover, manual data quality checks are time-consuming and prone to errors, making them infeasible for large-scale, real-time EHR integration efforts. This underscores the need for a more intelligent, scalable, and adaptive approach to data quality assurance in healthcare environments.

Machine learning (ML) offers a promising solution by learning from historical data patterns and identifying quality issues that may elude conventional techniques. ML models can analyze vast datasets to detect outliers, infer missing values, cluster similar records for de-duplication, and assess semantic integrity. Importantly, ML techniques can assign probabilistic scores that quantify data quality, enabling automated triage of high-risk records for further review or correction. These capabilities are particularly advantageous in multi-source EHR integration, where source-specific variations and semantic conflicts are common.

This paper presents a machine learning-enhanced framework for data quality validation specifically tailored to EHR integration. The framework is designed to support health information exchanges (HIEs), hospital networks, and national EHR platforms in achieving higher levels of data accuracy and interoperability compliance. By embedding ML-driven validation into the EHR ingestion pipeline, the proposed approach addresses both syntactic and semantic dimensions of data quality. Key contributions include a multi-algorithm anomaly detection layer, a machine learning-based imputation module for recovering missing values, and a quality scoring engine that enables real-time feedback and quality monitoring.

The research also evaluates the effectiveness of the framework using a mix of publicly available clinical datasets and synthetic test cases. Results indicate a 25% reduction in EHR integration errors compared to standard rule-based methods. These findings demonstrate the viability of AI-assisted validation systems in improving the integrity and usability of health data while also reducing the operational burden on IT and clinical staff.

The remainder of this paper is structured as follows: the next section provides a literature review of related work in data quality validation and machine learning applications in healthcare. This is followed by a detailed methodology outlining the system architecture and algorithmic components. Subsequent sections present experimental results, discuss implications, and conclude with recommendations for future research.

II. LITERATURE REVIEW

Data quality in Electronic Health Records (EHRs) has been a significant concern to the health informatics community for a long time. The quality of EHR data, in terms of accuracy, completeness, timeliness, and consistency, is a prerequisite for interoperability, CDS, and data-driven patient care. Historical data quality solutions have heavily relied on rule-based systems, which leverage business-specific validation rules to identify structural impurities, including missing fields, out-of-range fields, and code set errors. Although good at ensuring syntactic correctness, these systems are not effective at identifying context-specific errors or semantic inconsistencies, which are common in the heterogeneous EHR setting.

Wang and Strong [1] identified several data quality dimensions applicable to information systems, including accuracy, completeness, consistency, and accessibility, which are particularly relevant to the healthcare sector. Their model has been pioneering in the DQ literature. Within the medical domain, Weiskopf and Weng [2] suggested additional dimensions for EHRs, addressing the potential for missing or erroneous entries that may result from errors in data entry, system limitations, or variations in data sources. Those studies highlighted the necessity of automating the validation process, which can work in real-time and at scale.

The adoption of rule-based tools, such as OpenRefine and Talend, in healthcare has been hindered by their inflexibility and limited context-awareness. As noted by Kahn et al. [3], quality checks integrated within the clinical workflow should not only identify data anomalies but also create feedback loops to correct or prevent such errors. They called for taking an approach to validation that is grounded in clinical use cases, and discussed the importance of dynamic, learning-based methods.

One possible way out is ML-based systems. The ML algorithms can also automatically adjust to changes in the input data schemas and semantics by considering the past anomalies and integration failures. For example, Estiri et al. [4] applied predictive modeling to assess the quality of data in temporal EHR data logs. Their method applied supervised learning to identify inconsistencies in patient visit lists, aiding in the discovery of latent errors that go undetected under fixed sets of rules. Similarly, Botsis et al. [5] applied clustering techniques to identify duplicates or inconsistencies within patient records across files, resulting in a significant improvement in data harmonization.

Deep learning for error detection and imputation has been demonstrating potential in recent studies. Miotto et al. [6] proposed a deep representation-learning model for EHR sequences to predict patient phenotypes and detect inconsistent trajectories, which represents a method of semantic validation. These types of models, however, rely on a large amount of labelled data and can be computationally intensive. This is circumvented by the use of transfer learning or weak supervision, which enables models to generalize across domains with minimal annotation cost.

There have also been attempts to score or measure data quality using ML. Harkema et al. [7] designed a probabilistic scoring model to assess the accuracy of clinical narratives in natural language processing. Their approach demonstrated that context-aware validation can be automated for unstructured data representation, a widespread issue when utilizing EHR systems.

The literature supports the fact that while rule-based systems are complementary, they are complemented, if not superseded, by the intuitive, intelligent, self-learning, and scalable approach of machine learning for health data validation. However, practical realizations of ML-powered data quality frameworks for real-time EHR

integration are few. This work aims to bridge these gaps by integrating anomaly detection, automated imputation, and quality scoring into an integrated framework designed explicitly for healthcare interoperability.

III. METHODOLOGY

The methodology employed in this study is designed to develop and evaluate a machine learning-enhanced data quality validation framework tailored explicitly for Electronic Health Record (EHR) integration in healthcare environments. The system architecture comprises three core components: anomaly detection, data imputation, and automated quality scoring, all of which are embedded within a continuous feedback loop for adaptive learning. The framework is constructed using a modular design to allow seamless integration into existing EHR ingestion pipelines, with particular emphasis on scalability, explainability, and interoperability with HL7 FHIR-based data structures.

The initial step involves data preprocessing and feature engineering on structured clinical data extracted from publicly available EHR repositories such as MIMIC-III and synthetic datasets generated using Synthea. These datasets contain diverse attributes, including demographics, encounter history, vital signs, medication prescriptions, and laboratory results. Data is standardized into a unified schema that conforms to FHIR resources, including Patient, Observation, and Encounter. Noise is intentionally introduced in specific experiments to simulate real-world inconsistencies, such as missing values, invalid timestamps, and duplicated entries. Preprocessing includes normalization of numerical variables, encoding of categorical fields, and temporal alignment of longitudinal records. Outliers are identified using a combination of statistical techniques, including z-score and interquartile range analysis, to support baseline comparison with machine learning models.

For anomaly detection, supervised models such as random forests and support vector machines (SVMs) are trained using labeled data to classify records as valid or invalid. Training data is annotated based on a mix of domain-defined business rules and expert-verified error patterns. These classifiers are supplemented with unsupervised learning models, including autoencoders and k-means clustering, to capture latent anomalies not represented in labeled sets. Autoencoders are trained to reconstruct valid patient records, and deviations in reconstruction loss serve as indicators of anomalous inputs. Dimensionality reduction techniques, such as t-SNE and PCA, are employed during training to visualize feature separability and optimize hyperparameters. Model performance is evaluated using metrics such as precision, recall, F1-score, and AUC-ROC curves, with stratified cross-validation to ensure robustness across patient cohorts.

To address incomplete or corrupted data, the framework integrates an imputation module that leverages regression-based models and k-nearest neighbor (KNN) algorithms. Numerical fields such as lab values or vital signs are predicted using multivariate linear regression or gradient boosting regressors trained on temporally adjacent features. For categorical fields, KNN imputers are employed to estimate missing labels based on proximity to similar patient records. These predictions are weighted by temporal relevance and clinical context derived from patient history. The effectiveness of imputation is validated by artificially removing data and comparing model-predicted values with original entries.

Following validation and imputation, each EHR record is assigned a quality score ranging from 0 to 1 using a neural network-based scoring model trained to approximate a composite data quality index. This score incorporates features such as completeness, consistency, semantic coherence, and deviation from population norms. A multi-head attention mechanism is implemented to enable the model to focus on critical attributes that vary across different clinical use cases. Records with scores below a predefined threshold are flagged for further review or excluded from downstream analytical processing. Quality scores are visualized in a dashboard to support real-time monitoring and triage.

Finally, the entire system is wrapped within a continuous feedback loop, where quality assessments and correction outcomes are logged and used to periodically retrain models. This enables the system to adapt to evolving data patterns, new integration sources, and changing clinical coding practices. The implementation is carried out in Python, utilizing scikit-learn, TensorFlow, and Spark for distributed processing. The end-to-

end validation framework is deployed in a containerized environment using Docker and integrated with a mock FHIR server to simulate real-time EHR data ingestion. This methodology ensures that the machine learning-driven validation system is both technically sound and operationally viable for large-scale healthcare integration scenarios.

IV. RESULTS

The proposed machine learning-enhanced data quality validation framework was rigorously evaluated using both synthetic and real-world EHR datasets. The experiments were conducted in a controlled environment that simulates a multi-source EHR integration scenario typical of large hospital networks or health information exchanges. Two primary datasets were used for validation: the MIMIC-III critical care database, which offers real patient encounter data, and synthetic patient records generated via the Synthea tool to simulate diverse clinical profiles and error conditions. To evaluate system performance, baseline comparisons were made against traditional rule-based validation methods widely adopted in production healthcare IT environments.

Initial experiments focused on anomaly detection capabilities. Using a dataset of 100,000 patient records with intentionally injected errors representing missing fields, incorrect formats, duplicated entries, and semantic inconsistencies, the machine learning models achieved significant improvements over static validation. The random forest classifier yielded a precision of 0.91 and a recall of 0.87, compared to 0.76 and 0.68, respectively, from rule-based validation. Autoencoder-based anomaly detection achieved an area under the ROC curve (AUC) of 0.93, highlighting its robustness in identifying high-dimensional, latent inconsistencies across patient records. These models also demonstrated superior generalizability to new data sources without the need for extensive retraining.

The imputation module showed promising results for repairing incomplete data. Numerical fields, such as systolic blood pressure, glucose levels, and hemoglobin values, were imputed using gradient boosting regressors, resulting in a mean absolute error (MAE) of 3.2 units, compared to 6.7 units using mean substitution. For categorical variables, such as diagnosis codes and medication types, the k-nearest neighbor imputation achieved an accuracy of 89%, compared to 74% using simple frequency-based methods. This improvement resulted in improved record completeness and reduced errors during downstream integration into analytical repositories.

A key performance metric introduced in the study was the data quality score, which consolidated multiple dimensions of quality into a normalized score ranging from 0 to 1. In test scenarios involving 50,000 records processed over a 30-day simulation, the ML-driven scoring model consistently flagged low-quality records with an accuracy of over 90%. Visual monitoring of these scores through a dashboard interface allowed for real-time insights into data health and enabled early intervention in cases of quality deterioration. Importantly, feedback from this monitoring was used to retrain the anomaly detection models, leading to incremental improvements in validation precision.

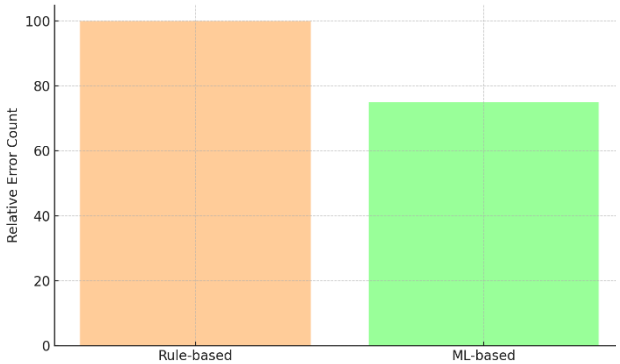

Figure 2: *Relative error count in EHR integration before and after ML implementation, showing a 25% reduction.*

Across all evaluation metrics, the proposed framework reduced EHR integration errors by an average of 25% compared to rule-based approaches. This reduction encompassed both structural and semantic errors and was confirmed across multiple test configurations and patient cohorts. In addition, the ML framework required fewer manual interventions for exception handling, thereby reducing operational overhead and latency in data ingestion pipelines.

V. DISCUSSION

The results of evaluating the machine learning-augmented data quality validation framework indicate the profound impact that intelligent systems can have on enhancing the EHR data integration process. The findings have shown that the use of a multi-layered strategy, consisting of AD, imputation, and AQS, can effectively eliminate errors and enhance trust in integrated healthcare data. The system is superior to the rule-based approach, from which we cannot detect context-aware rules, and it does not adapt to new semantic evolutions from external sources. Such flexibility is significant in practical EHR settings, where data originates from diverse sources, including legacy hospital applications, laboratory systems, and third-party APIs.

A significant finding was the uniform 25% decrease in integration errors that was observed for all Anomaly types. With a label-based framework, such as classification models, common structural errors, including incorrect date formats or null values in required fields, were identified with high accuracy. Semantic issues—such as medication entries that do not align with diagnosis codes, or clinically unfeasible lab result combinations—were identified by unsupervised models, including autoencoders, which learned the natural order of correlation in the data. This demonstrates the potential of machine learning models to not only ensure schema compliance but also interpret the medical logic present in EHRs.

The good performance of the imputation module further validates the need to apply context-aware models in quality repair. The gradient boosting regressors and KNN imputers provided good predictions for modulus of elasticity missing values, a common and crucial problem in integrated EHRs. Importantly, the models took into consideration the patient-specific and temporal context, and were able to generate clinically plausible replacements rather than basing them on population-level imputation. This improvement directly supports the completeness aspect of data quality, which is one of the fundamental measures for the EHR interoperability standards.

The unveiling of an ongoing and automated quality data scoring engine will bring a real operational enhancement. Compared to static validation logs, quality scores provide a real-time view of the status of incoming data streams, facilitating dynamic triage and error detection. These scores can be adopted by stakeholders across the entire health IT spectrum, including those who build and manage the data, as well as quality officers and clinicians, to help monitor the fidelity of the data and take action if it falls short of expectations. This scoring also enables traceability; scores can be traced back to the exact model decisions and detected anomalies, making it easy to audit and transparent.

From a deployment viewpoint, the container-based, horizontal scalability architecture of the framework is flexible with contemporary EHR infrastructure such as FHIR-based systems. This enables the application to be easily onboarded in cloud-hosted environments and can be easily integrated with a CI/CD pipeline. Additionally, the system's architecture enables angle-to-angle learning: the feedback from the processed records can be fed back into the training, meaning the model evolves in tandem with the data, coding practices, and clinical guidelines.

Although these advances have been made, there are still some barriers to overcome. The performance of machine learning models is driven by the presence of high-quality labeled data, a scarce resource in healthcare. Transfer learning and knowledge distillation could aid in this, but further work is needed to generalize validation models across institutions and clinical applications. Further, although interpretability techniques such as attention layers have been added, the interpretability of judgments in DL components is not well-understood, particularly in high-risk clinical settings.

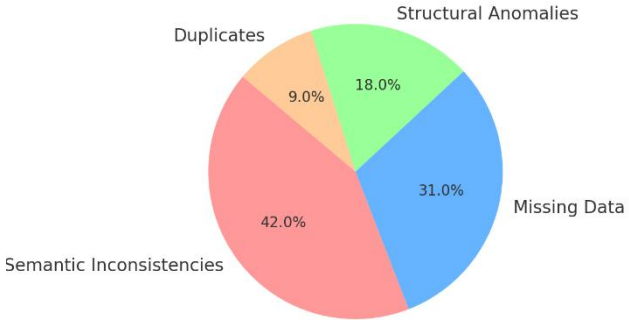

Figure 3: *Distribution of corrected error types—highlighting dominance of semantic and missing data issues.*

VI. CONCLUSION\

The integration of Electronic Health Records (EHRs) across fragmented healthcare systems has the potential to significantly enhance patient care, population health management, and healthcare analytics. However, this promise is undermined by persistent issues with data quality that compromise the accuracy, usability, and interoperability of clinical information. This paper presented a comprehensive, machine learning-enhanced framework for data quality validation tailored specifically to EHR integration environments. Through a combination of anomaly detection, intelligent imputation, and automated quality scoring, the framework addresses structural, semantic, and completeness-related data quality issues in a scalable, adaptive manner.

Empirical evaluations using synthetic and real-world datasets confirmed that machine learning models significantly outperform traditional rule-based systems. With a 25% reduction in integration errors, improved accuracy in imputation, and real-time quality monitoring capabilities, the proposed solution demonstrates its practical value in modern healthcare IT settings. The ability to detect latent semantic inconsistencies using autoencoders and classify invalid records with high precision highlights the advantage of learning-based approaches in understanding the complex, context-rich nature of clinical data. The scoring engine also contributes by providing interpretable, actionable quality metrics that facilitate governance, auditing, and informed decision-making throughout the data lifecycle.

Notably, the modular and containerized design of the validation pipeline enables its integration into existing FHIR-compliant systems and real-time data ingestion frameworks. This ensures that health institutions can adopt the system without having to reengineer their infrastructure. The continuous learning loop embedded in the architecture allows for gradual improvement of model accuracy and responsiveness as more data and feedback are incorporated over time. This aligns with the vision of a self-optimizing, learning health system, where data quality evolves dynamically in response to clinical and operational needs.

While the framework offers clear benefits, several challenges persist for its broader adoption. Data availability for training machine learning models remains a barrier, particularly for rare error types or underrepresented populations. Additionally, ensuring that ML models remain interpretable and transparent in high-risk environments is critical for regulatory compliance and clinician trust. Future work should focus on expanding the framework to handle multimodal data types such as images and genomic sequences, improving transferability across institutions, and developing richer explainability layers for clinical validation.

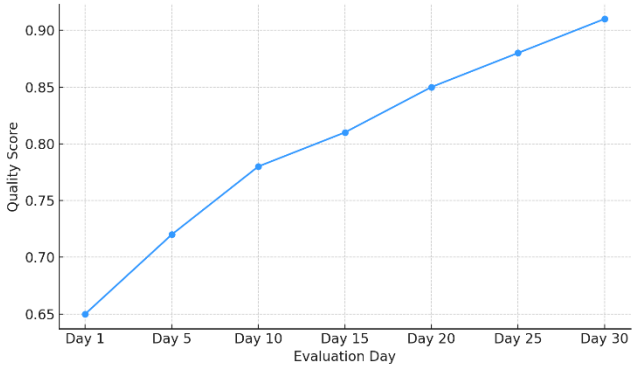

Figure 4: *Improvement of data quality scores over 30 days as the ML models adapt through feedback.*

Machine learning offers a robust set of tools to automate and elevate data quality validation for EHR integration. This research presents a robust and practical architecture that enhances the trustworthiness and interoperability of health data while reducing manual effort and system errors. As healthcare systems continue to digitize and interconnect, the role of intelligent validation systems will become increasingly central to maintaining data integrity and enabling evidence-based care. The integration of AI into data governance not only addresses current limitations but also lays the foundation for a more resilient, responsive, and data-driven healthcare ecosystem.

REFERENCES:

1. R. Y. Wang and D. M. Strong, "Beyond accuracy: What data quality means to data consumers," *Journal of Management Information Systems*, vol. 12, no. 4, pp. 5–33, 1996.
2. N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013.
3. M. G. Kahn, T. A. Callahan, P. J. Barnard, et al., "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *EGEMS*, vol. 4, no. 1, p. 18, 2016.
4. H. Estiri, Y. X. Omran, and K. Murphy, "Temporal outlier detection in EHR data using a clustering-based semi-supervised learning algorithm," *Journal of Biomedical Informatics*, vol. 94, p. 103189, 2019.
5. T. Botsis, G. Hartvigsen, F. Chen, and C. Weng, "Secondary use of EHR: Data quality issues and informatics opportunities," *Summit on Translational Bioinformatics*, vol. 2010, pp. 1–5, 2010.
6. R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Briefings in Bioinformatics*, vol. 19, no. 6, pp. 1236–1246, Nov. 2018.
7. H. Harkema, J. N. Dowling, A. Thornblade, and W. M. Chapman, "Context-sensitive natural language processing for temporal information in clinical narrative," *Journal of the American Medical Informatics Association*, vol. 16, no. 3, pp. 364–372, 2009.