iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77218

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

Research Analysis for Detecting Duplicate Entries in Electoral Voter List

Krish Pritesh Keer, Prof. Suhas Rautmare (Supervisor)
Department of Information Technology University of Mumbai, Kalina

Abstract: ***Electoral voter lists form the backbone of democratic processes, ensuring that every eligible citizen is granted a fair opportunity to vote. However, large-scale voter databases often suffer from data quality issues such as duplicate or near-duplicate entries caused by spelling variations, data entry errors, migration, and inconsistent demographic updates. These redundancies can compromise the integrity of elections, increase administrative costs, and reduce public trust. This research proposes an intelligent, analytics-driven framework for detecting duplicate entries in electoral voter lists using a combination of data pre processing techniques, rule-based matching, and machine learning models. The proposed system integrates phonetic similarity, demographic attribute comparison, and supervised classification models to identify potential duplicates with high accuracy. Experimental analysis demonstrates that the hybrid approach significantly outperforms traditional exact-matching techniques, offering a scalable and reliable solution for election management bodies. The study emphasizes transparency, accuracy, and scalability while maintaining compliance with ethical and data privacy considerations.***
Keywords: Electoral voter list, duplicate detection, data analytics, record linkage, data quality.

I. INTRODUCTION

Electoral voter lists are critical national databases that ensure the legitimacy, transparency, and fairness of democratic elections. An accurate voter registry guarantees that every eligible citizen can exercise their voting rights while preventing irregularities such as redundant registrations, administrative inefficiencies, and misallocation of electoral resources. In large democracies, voter databases may contain millions of records collected over long periods, across different regions, and by multiple authorities, making them inherently vulnerable to data quality issues.

Duplicate entries in electoral voter lists commonly arise due to spelling inconsistencies in names, variations in address formats, migration of voters between constituencies, delayed deletion of outdated records, and manual data entry errors. In multilingual and culturally diverse societies, the same individual's name may be recorded differently across regions, further increasing the complexity of duplication detection. These challenges make traditional exact-matching techniques insufficient and unreliable.

Manual verification of voter lists is time- consuming, costly, and prone to human error. As electoral databases grow in scale, the need for automated, intelligent systems becomes increasingly urgent. Advances in data analytics, artificial intelligence, and machine learning offer robust solutions for identifying hidden patterns and similarities within large datasets. By learning from historical data and leveraging similarity measures machine learning models can effectively detect both exact and near-duplicate records.

This research aims to design and analyze an intelligent analytical framework for detecting duplicate entries in electoral voter lists. The proposed approach combines preprocessing techniques, similarity-based feature engineering, and supervised machine learning algorithms to enhance accuracy and scalability. The study contributes to the field by addressing real-world data challenges while maintaining transparency, fairness, and ethical considerations in electoral data management.

The presence of duplicate voter records does not necessarily indicate intentional misuse but often reflects systemic limitations in large-scale data collection and maintenance processes. Factors such as decentralized registration procedures, delayed synchronization between regional offices, and differences in data recording standards contribute significantly to duplication. Addressing these challenges requires automated mechanisms capable of understanding contextual similarities rather than relying solely on exact attribute matching.

An intelligent analytical approach enables election authorities to proactively identify redundant entries, streamline database maintenance, and improve operational efficiency. By adopting data-driven techniques, electoral systems can shift from reactive manual corrections to predictive and preventive data quality management. This study focuses on developing such an approach using analytical modeling and machine learning techniques tailored specifically for electoral voter data.

Conventional approaches for duplicate detection rely heavily on deterministic rules or manual verification processes. While such methods are simple to implement, they lack scalability and are unable to detect near-duplicate records that differ marginally across attributes. Manual verification, in particular, is time-consuming, costly, and susceptible to human error. These limitations highlight the need for automated and intelligent solutions capable of processing large volumes of electoral data efficiently.

Recent advancements in data analytics and machine learning have enabled the development of sophisticated methods for identifying hidden patterns and similarities within large datasets. By leveraging similarity metrics and supervised learning algorithms, it is possible to detect both exact and approximate duplicates with higher accuracy. This research focuses on designing an intelligent framework that applies these techniques to electoral voter lists, aiming to improve data quality while ensuring transparency, fairness, and ethical compliance.

II. LITERATURE REVIEW

The problem of duplicate detection, also known as record linkage or entity resolution, has been widely studied across multiple domains such as census data management, healthcare information systems, banking, and customer databases. Early research in this area focused on deterministic or rule-based methods, where records were considered duplicates if specific attributes matched exactly. While computationally efficient, these methods fail in real-world scenarios where data is incomplete, noisy, or inconsistent.

Fellegi and Sunter introduced a probabilistic framework for record linkage that assigns weights to matching and non-matching attributes, enabling partial matches to contribute to an overall similarity score. This approach improved recall and flexibility but required extensive domain knowledge and manual tuning. Subsequent studies explored heuristic-based methods using string similarity metrics such as Levenshtein distance, Jaro-Winkler similarity, and token-based matching to handle typographical variations.

With the evolution of machine learning, researchers began applying supervised and unsupervised algorithms to automate duplicate detection.

Decision trees, support vector machines, random forests, and gradient boosting algorithms demonstrated strong performance in learning complex relationships between record attributes. Recent studies also explored clustering and deep learning-based embedding techniques for large- scale entity resolution tasks.

In the electoral domain, research efforts have focused on phonetic matching of names, demographic attribute comparison, and geographic clustering to identify duplicate voter registrations.

However, many existing approaches struggle with scalability, lack adaptability to regional naming conventions, or rely heavily on manual verification. This research addresses these gaps by proposing a hybrid framework that integrates domain-specific rules with machine learning models to achieve both accuracy and scalability in electoral voter list management.

Duplicate detection, also known as record linkage or entity resolution, has been an active area of research across domains such as census management, healthcare systems, financial databases, and customer relationship management. Early research primarily focused on deterministic methods, where records were classified as duplicates based on exact matches across selected attributes. Although computationally efficient, these approaches are unsuitable for real-world datasets characterized by noise, missing values, and inconsistent formatting.

Probabilistic record linkage models marked a significant advancement by assigning weights to matching and non-matching attributes and computing an overall similarity score. These models improved flexibility and recall but required careful parameter tuning and domain expertise.

Subsequent research explored heuristic-based techniques using string similarity metrics, such as edit distance and token-based matching, to handle typographical variations and partial matches.

With the emergence of machine learning, researchers began applying supervised and unsupervised algorithms to automate duplicate detection. Classification models such as decision trees, random forests, and gradient boosting demonstrated strong performance in learning complex relationships between record attributes. More recent studies have investigated deep learning approaches and embedding-based similarity representations to address large-scale entity resolution challenges.

Recent research trends emphasize learning-based approaches that automatically infer similarity patterns from data. Supervised learning models have shown particular promise in domains where labeled examples are available, enabling systems to distinguish meaningful variations from genuine mismatches. In parallel, hybrid models that integrate domain-specific heuristics with machine learning have emerged as effective solutions for large-scale record reconciliation tasks.

Extensive manual validation. This research builds upon existing literature by proposing a hybrid framework that integrates domain knowledge with machine learning to achieve both accuracy and operational efficiency in electoral voter list management.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

III. PROBLEM STATEMENT

Despite advances in digital record management, electoral voter lists continue to face challenges related to data redundancy and inconsistency.

Existing systems primarily rely on static rules or manual verification processes, which are insufficient for detecting near-duplicate records in large datasets. There is a lack of scalable, intelligent mechanisms capable of identifying duplicates while accommodating variations in voter information.

The problem addressed in this research is the absence of an automated, accurate, and adaptable framework for detecting duplicate voter entries that can operate efficiently on large-scale electoral databases while maintaining transparency and administrative control.

IV. METHODOLOGY

The proposed system follows a structured, multi- stage pipeline designed to efficiently identify duplicate entries within large-scale electoral voter databases. Each stage of the framework addresses specific challenges associated with data quality, scalability, and accuracy.

A. Data Collection and Preprocessing

The system assumes access to structured voter registration data containing attributes such as voter name, date of birth, age, gender, address, constituency, and identification numbers.

Preprocessing is a critical step to improve data consistency and reliability. It includes:

1) Removal of duplicate headers and irrelevant fields
2) Elimination of incomplete or invalid records
3) Standardization of text fields through case normalization
4) Removal of special characters, punctuation, and extra spaces
5) Expansion of common abbreviations in names and addresses

B. Feature Engineering

Effective duplicate detection depends on extracting meaningful similarity features from raw data. The proposed system generates multiple features, including:

1) Phonetic Similarity: Encodes names using phonetic algorithms to capture sound-based similarities
2) String Similarity Metrics: Calculates edit distance and token overlap for names and addresses
3) Demographic Consistency: Compares date of birth, age difference, and gender
4) Geographic Similarity: Evaluates proximity based on address, locality, or constituency

C. Blocking Strategy

To reduce the computational complexity of pairwise comparisons, blocking techniques are applied. Records are grouped into smaller blocks using attributes such as postal code, constituency, or birth year. Only records within the same block are compared, significantly improving processing efficiency without sacrificing detection accuracy.

D. Decision Threshold and Validation

A configurable threshold is applied to the predicted probability scores to classify record pairs. High-confidence duplicates may be automatically flagged, while borderline cases are forwarded for human verification. This human-in-the-loop approach ensures transparency, accountability, and continuous improvement of the system. The proposed duplicate detection framework follows a structured, multi-stage pipeline designed to efficiently process large-scale electoral voter databases. The system is modular in nature, allowing each stage to operate independently while maintaining a continuous flow of information across components.

E. Data Collection and Preprocessing

The framework assumes access to structured voter registration data containing attributes such as voter name, date of birth, gender, address, and constituency information. Preprocessing involves standardizing textual fields, removing invalid or incomplete records, handling missing values, and normalizing formats. These steps reduce noise and improve consistency, thereby enhancing the effectiveness of subsequent similarity analysis.

F. Feature Engineering and Similarity Computation

Multiple similarity features are generated to capture variations across voter attributes. Phonetic encoding techniques are applied to names to address spelling differences, while string similarity metrics quantify textual variations in names and addresses. Demographic attributes are compared to ensure consistency, and geographic information is analyzed to identify spatial proximity. Each feature contributes to an aggregated similarity score representing the likelihood of duplication.

G. Blocking and Pair Generation

To improve scalability, blocking strategies are employed to limit comparisons to records within the same group, such as shared locality or birth year. This approach significantly reduces computational complexity while preserving detection accuracy.

H. Machine Learning Classification

Supervised machine learning models are trained on labeled record pairs to distinguish between duplicate and non-duplicate entries. Ensemble- based classifiers are preferred due to their robustness and ability to handle heterogeneous features. The model outputs probability scores, enabling flexible threshold-based decision-making.

I. Decision Validation

High-confidence duplicates are flagged for resolution, while ambiguous cases are forwarded for human review. This human-in-the-loop mechanism ensures accountability, transparency, and continuous improvement of the system.

Data Blocking and Candidate Pair Reduction One of the primary challenges in duplicate detection within large electoral voter databases is the computational cost associated with comparing every record with all others.

A naïve pairwise comparison approach results in quadratic time complexity, which is infeasible for datasets containing millions of voter records. To address this challenge, the proposed framework incorporates a blocking mechanism that significantly reduces the number of candidate record pairs.

Blocking involves grouping records into smaller subsets based on shared attribute values such as locality, postal code, or year of birth. Only records within the same block are compared further for duplication.

This strategy maintains high detection accuracy while drastically reducing processing time. Multiple blocking keys can be combined to improve coverage and minimize missed duplicates. The use of blocking ensures that the system remains scalable and suitable for real-world electoral databases.

Handling Missing, Inconsistent, and Noisy Data Electoral voter data often contains missing values, inconsistent formats, and noisy entries due to manual data entry and periodic updates. The proposed system addresses these issues through a robust data handling strategy.

Missing values are treated using attribute- specific techniques such as imputation, default placeholders, or exclusion when necessary.

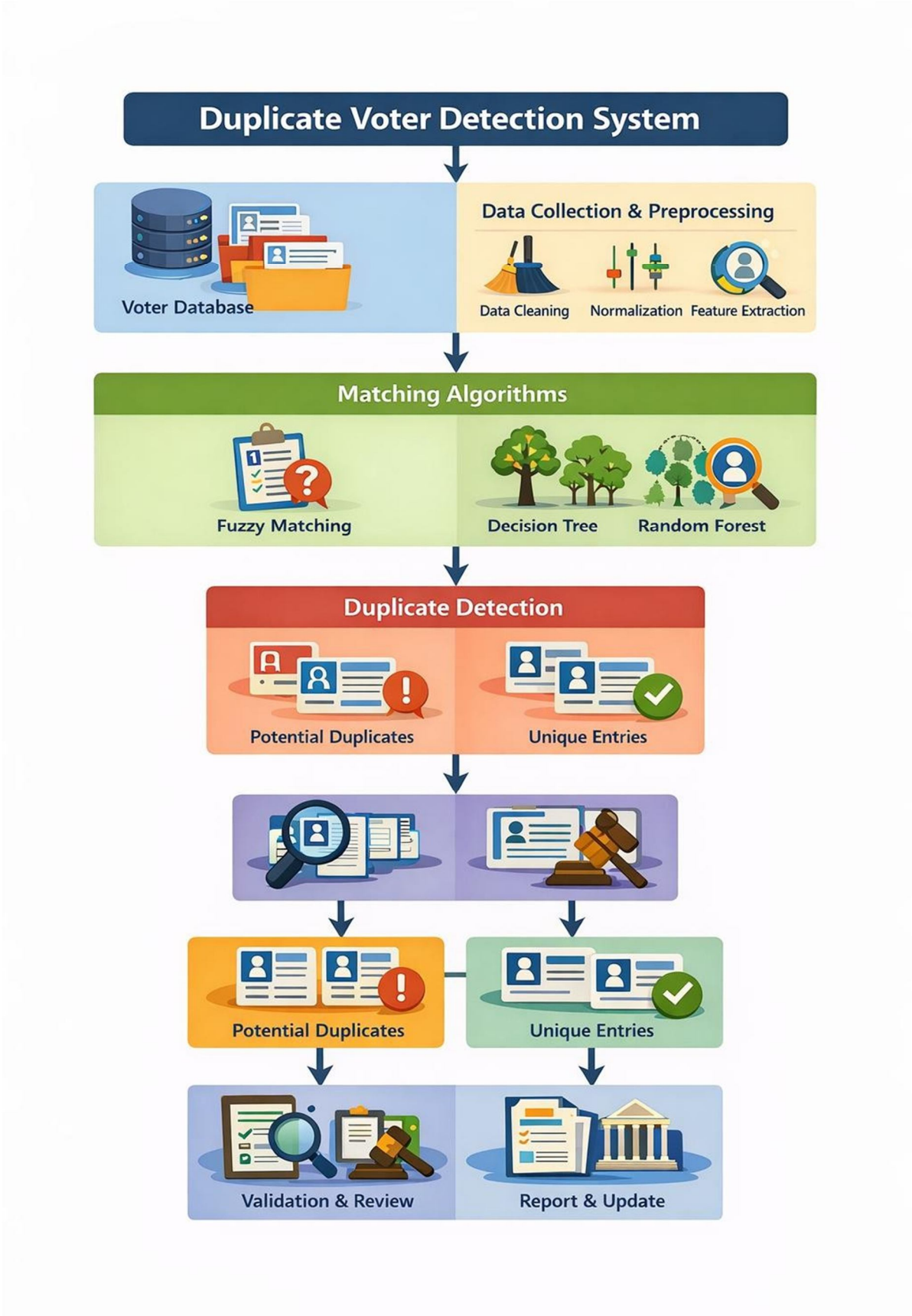
Inconsistent formats, particularly in textual attributes like names and addresses, are standardized using normalization rules.

Noise caused by abbreviations, typographical errors, or alternate spellings is mitigated through fuzzy matching and phonetic encoding. By addressing data imperfections at an early stage, the system ensures reliable similarity computation and reduces false detection rates. This preprocessing robustness is essential for maintaining accuracy in large and diverse voter datasets.

Threshold Selection and Decision Logic

The machine learning classifier generates probability scores indicating the likelihood that a pair of voter records represents the same individual. Determining an appropriate decision threshold is critical to balancing false positives and false negatives. In the proposed framework, threshold values are configurable and can be adjusted based on operational requirements.

High-confidence duplicates that exceed the upper threshold are automatically flagged for resolution, while pairs below the lower threshold are classified as unique. Intermediate cases are marked for manual verification. This tiered decision logic ensures flexibility and supports domain-specific requirements of electoral authorities. The ability to fine-tune thresholds enhances both precision and recall, depending on administrative priorities.



V. DISCUSSION

The proposed analytical framework demonstrates significant improvements over traditional duplicate detection techniques used in electoral voter list management. By integrating rule-based preprocessing with machine learning classification, the system effectively balances interpretability and predictive performance. Phonetic and string similarity features play a crucial role in identifying duplicates caused by spelling variations and regional naming differences.

The use of blocking strategies enhances scalability, allowing the framework to process large datasets efficiently. Ensemble learning models provide stable and accurate predictions, even in the presence of noisy or incomplete data. Furthermore, the modular design of the framework allows election authorities to adapt the system to local requirements and data formats.

Ethical considerations are central to the proposed approach. The system avoids the use of sensitive or unnecessary attributes and supports transparent decision-making through probability-based outputs. Human oversight in final decision-making helps mitigate algorithmic bias and ensures compliance with data protection regulations. The proposed framework demonstrates clear advantages over traditional duplicate detection techniques. By integrating similarity-based feature engineering with machine learning classification, the system effectively handles data inconsistencies and near- duplicate cases that deterministic methods fail to detect. Blocking strategies ensure scalability, enabling the framework to process large voter databases efficiently.

The use of probability-based outputs enhances interpretability and supports informed decision- making by election authorities. Ethical considerations are central to the framework's design, with a focus on minimizing bias and avoiding unnecessary use of sensitive personal information. Human oversight further strengthens trust and accountability in automated decision processes.

VI. CONCLUSION AND FUTURE SCOPE

This research presents an intelligent and scalable framework for detecting duplicate entries in electoral voter lists using data analytics and machine learning techniques. The proposed system effectively addresses the limitations of traditional rule-based approaches by handling data inconsistencies, spelling variations, and large-scale datasets. The integration of similarity-based features and supervised learning models enhances detection accuracy while maintaining operational efficiency.

The findings suggest that the proposed framework can significantly support election management bodies in maintaining clean and reliable voter databases. By reducing redundancy and improving data quality, the system contributes to administrative efficiency and public trust in electoral processes.

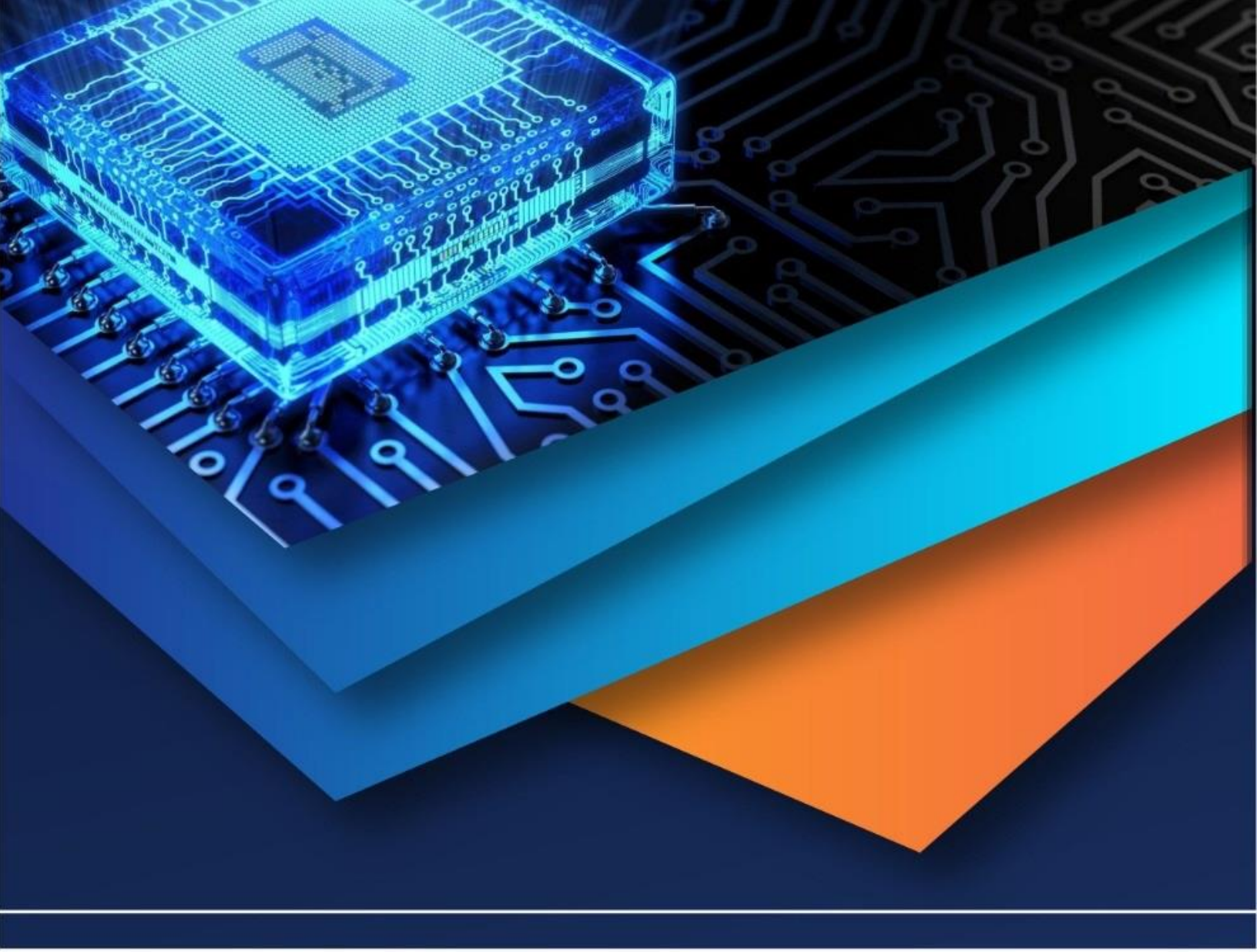
Future research may focus on incorporating deep learning-based similarity representations, multilingual name handling, and real-time voter list updates. Additionally, extending the framework to include explainable AI techniques can further enhance transparency and acceptance among stakeholders. Integration with national identity systems, while maintaining strict privacy safeguards, also presents a promising direction for future work.

Future enhancements may include the incorporation of multilingual text processing to handle regional language variations, adaptive learning mechanisms that update models based on new data patterns, and integration with real-time voter registration platforms. Expanding explainability features will further support trust and transparency in automated electoral data management systems.

Future work may explore the integration of deep learning-based similarity models, multilingual processing capabilities, and real-time duplicate detection during voter registration. Additionally, incorporating explainable artificial intelligence techniques can further enhance transparency and acceptance among stakeholders. The proposed framework provides a strong foundation for developing next-generation electoral data management systems that are accurate, ethical, and scalable.

REFERENCES

[1] Christen, P. (2012). Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer.
[2] Elmagarmid, A. K., Ipeirotis, P. G., & Verykios, V. S. (2007). Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering.
[3] Winkler, W. E. (2006). Overview of record linkage and current research directions. Bureau of the Census.
[4] Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association.
[5] Bhattacharya, I., & Getoor, L. (2007).
[6] Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data.
[7] Dalvi, N., Kumar, R., & Soliman, M. (2012). Automatic record linkage. Proceedings of the VLDB Endowment.
[8] Hernández, M. A., & Stolfo, S. J. (1998). Real- world data is dirty. Data Mining and Knowledge Discovery.
[9] Rahm, E., & Do, H. H. (2000). Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin.
[10] Köpcke, H., & Rahm, E. (2010). Frameworks for entity matching. Data & Knowledge Engineering.
[11] Getoor, L., & Machanavajjhala, A. (2012). Entity resolution: Theory, practice & open challenges. Proceedings of the VLDB Endowment.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY