Dijlah Journal of Engineering Science (DJES) Vol. 3, No. 1, March, 2026, pp. 12-30
ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 106

Hybrid Big Data Architecture Using SQL, NoSQL, and Machine Learning: A Case Study on Classifying Electric Vehicle Types

Ahmed Ibrahim Sharqi
Department of Informatics, Faculty of Engineering, Islamic University of Lebanon, Beirut, Lebanon

Article Info

Article history:

Received Sept., 26, 2025
Revised Dec. 20, 2025
Accepted Feb., 15, 2026

Keywords:

Hybrid Big Data
SQL
NoSQL
Machine Learning
Electric Vehicle Classification
Random Forest.

ABSTRACT

This study presents a robust hybrid big data architecture that integrates both relational (SQL) and non-relational (NoSQL) database systems to effectively manage and classify diverse electric vehicle (EV) data. The suggested method makes use of each database type's advantages enables efficient storage, retrieval, and processing of heterogeneous datasets. The data was preprocessed and evaluated using data mining techniques. followed by the implementation of multiple machine learning techniques, such as AdaBoost, Random Forest, K-Nearest Neighbours, Decision Tree, Support Vector Machine, and Logistic Regression —for accurate EV type classification. Among these, Random Forest demonstrated superior execution with a precision of 99.99%. The trained model was deployed through a real-time, user-friendly web interface to facilitate practical application and accessibility. This approach highlights the advantages of hybrid data architectures and machine learning integration, providing a scalable and adaptable framework applicable to other domains requiring heterogeneous data management and intelligent classification.

Corresponding Author:

Ahmed Ibrahim Sharqi
Department of Informatics, Faculty of Engineering, Islamic University of Lebanon, Beirut, LebanonAlmasafi street, Baghdad, Iraq
Email: ahmedibrahimsharqi@gmail.com

1. INTRODUCTION

The rapid expansion of the global economy and the depletion of natural resources have led to issues with sustainable development in a number of nations due to the energy crisis and environmental degradation [1,2]. As Figure 1 illustrates, one of the primary reasons for the high energy consumption and carbon emissions is the transport sector's resilience (data source: https://www.carbonmonitor.org.cn, retrieved on 6 July 2022) [3]. Consequently, nations have implemented diverse approaches to address the issues of pollution and the energy crisis. Because EVs save energy and reduce emissions and pollution, they are widely supported in the transportation sector [4-6]. Global sustainable development can be promoted and carbon emissions might be considerably decreased by converting from conventional to electric vehicles (EVs) fuel vehicles [7]. Concerns about environmental quality are addressed by the use of electric vehicles (EV), which also lessens reliance on internal combustion conventional vehicles (ICEV) [8]. Numerous studies view EVs as green products [[9–11]], which are products with minimal environmental impact [12], and as a possible way to reduce tailpipe air pollution [13] caused by ICEVs burning fossil fuels. When powered by renewable energy, EVs have the potential to significantly lower transportation-related GHG emissions [14]. EVs are more advantageous than ICEVs in regions with low-carbon power plants [15,16]. Although the EV industry has seen a significant increase in production over the past ten years, government policy support is likely to be needed in order to move towards a large-scale energy transition and achieve public goods from EVs. States, manufacturers, and consumers still require policies to promote the use of EVs

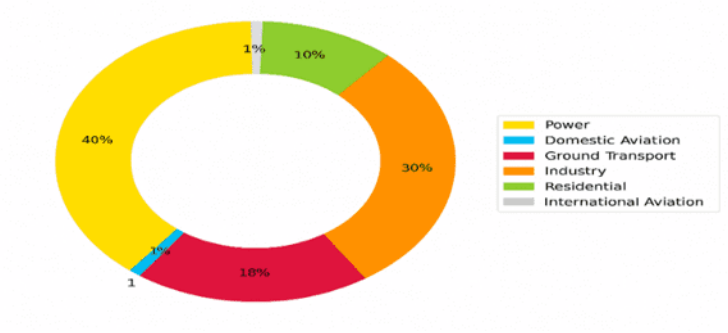

Figure1. Sector-specific carbon emissions in 2022

Between 2012 and 2020, EV sales in the US increased significantly [18,19]. Over 195,580 EVs were sold annually in 2017, a fourfold increase over 2012 sales figures [20]. Between 2010 and 2021, almost 2 million EVs were sold in the US [18]. Even though EV sales still make up a very small percentage of the automotive market—1.2% of all vehicles sold in 2017 [21]—2020 will see a notable increase that is consistent with the global EV market, as the U.S. EV market has grown to 1.8% of all vehicles [22]. In the United States, EV sales rose by 55% in 2022, accounting for 8% of total sales [23]. A number of nations have set goals to phase out ICEV as they transition to electric vehicles, including France, China, Norway, Ireland, and others [24, 25]. However, the growing population of electric vehicles (EVs) leads to the accumulation of vast and diverse datasets related to vehicle characteristics, registration records, manufacturer details, and usage patterns. These datasets often consist of structured data such as vehicle specifications and registration info, alongside semi-structured or unstructured data like textual descriptions and metadata. Efficiently managing and analyzing this heterogeneous, high-volume data is essential for accurately classifying EV types, understanding market trends, and supporting decision-making for policymakers, manufacturers, and energy providers. This growing influx of data is further driven by rapid industrialization, the increasing affordability of smart devices, considering the extensive use of innovative the Internet of Things, for example, cloud and edge computing, and ubiquitous mobile connectivity. The interconnection of billions of devices across various domains has given rise to an immense digital ecosystem frequently known as "big data" [26] [27]. Big data is characterized by its massive volume (ranging from terabytes to petabytes), various data types of data, including semi-structured, structured, and unstructured and high velocity — which together make real-time processing and analysis both critical and challenging. Due to the diverse and large-scale nature of electric vehicle population data, conventional relational database systems (RDBMS) often face challenges in efficiently storing, managing, and analyzing such datasets. Their limitations in scalability and flexibility have led to the increasing adoption of hybrid data management solutions that combine relational and non-relational databases. These big data technologies offer revolutionary potential in a variety of domains, such as market research, energy management, and transportation analytics [28]. Intelligent machine learning frameworks and sophisticated storage architectures are necessary to effectively classify and evaluate EV types by managing the amount and complexity of heterogeneous vehicle data at scale. Because of their strong querying capabilities, transactional integrity, and data consistency, relational database management systems (RDBMS), like SQL-based platforms, have long served as the basis for storing structured data about the population of electric vehicles. However, the shortcomings of conventional relational databases have emerged with the emergence of big data is defined by its large volume velocity, and variety. Semi-structured or unstructured data, such as textual descriptions, metadata, and sensor-generated data, are commonly found in EV population datasets but were not designed with these systems' effective management in mind. Additionally, they don't have the adaptability and horizontal scalability needed by contemporary, data-intensive applications that categorize and assess various EV types. [29] [30]. Consequently, researchers and businesses are increasingly choosing NoSQL (Not Only SQL) databases for big data environments. Distributed architectures, horizontal scalability, and schema flexibility are features offered by NoSQL systems such as Couchbase, Cassandra, and MongoDB. They are perfect for organizing and preserving the different kinds of data, such as sensor-generated data, textual descriptions, structured records, and metadata, that can be found in datasets pertaining to the population of electric vehicles. The complexity and diversity of EV data needed for precise classification and insightful analysis can be handled by NoSQL databases in addition to conventional SQL systems

[31]. NoSQL databases do have some drawbacks, even though they address many scalability issues and flexibility challenges associated with big data. A large number of NoSQL systems don't have or have reduced versions of RDBMS capabilities, such as full data consistency, efficient queries, and integrated transactions. A well-balanced hybrid data management solution is required to position electric vehicle population data to support appropriate classification and accurate analysis. The hybrid database model can now overcome the disadvantages of both SQL and NoSQL. The hybrid model combines NoSQL's speed and scalability with the advantages of relational databases, such as data integrity and ACID transactions. In a challenging and complex situation like population metrics of electric vehicle data, using a hybrid database model can assist with better real-time decisions, more accurate trend predictions, advanced analytics, and more precise classifications along with machine learning possibilities. Several contributions are made by this study to This hybrid model does more than just make the system more responsive and good access to data but also makes a complex system with big datasets easier to utilize resources smartly and in a scalable way, whether for businesses' policy decisions or understanding adoption trends. In order to access and analyze data about the population of electric vehicles, we created a hybrid big data architecture that incorporates the advantages of different database technologies, such as non-relational (NoSQL) and relational (SQL). The system utilizes SQL based databases such as MySQL or PostgreSQL to store structured data. This includes ownership information, original manufacturing data, and vehicle registration data. In addition, the system utilizes NoSQL databases such as MongoDB to store unstructured or semi-structured data, including metadata, textual descriptions, and external sources such as location wisdom or policy annotations. Machine learning models have been incorporated into the system for analytical purposes and allow for classification. Specifically, a model is being trained to classify the type of electric vehicle based on many factors (ex: model year, make, vehicle class, fuel type, etc.). The hybrid data architecture allows for easy access and preparation of data for training and evaluating classification models like (RF), (DT), or (LR). The purpose of classification models is to inform understanding of EV population patterns and to have predictive insight on the distributions of EV types that are possible. The meteoric rise in EV adoption has resulted in tidal waves of interesting population data collected on vehicles, including types, fuel types, years of manufacturer, manufacturers, and geographic distributions. Organizing, storing, and analyzing this population data is crucial needed for identifying patterns of use, monitoring market growth trends, and, importantly, for providing insights to inform business/policy decisions. In investigating novel ways to deal with these issues, researchers have explored a variety of data management strategies, including hybrid techniques that capitalize on the strengths between conventional relational databases (SQL) and non-relational databases (NoSQL). At the same time, machine learning techniques have attracted interest because of their ability to support automated classification of large EV datasets, predictive data modeling, and data-driven insights.

- **Data Storage Models (SQL, NoSQL, Hybrid)**

As the use of electric vehicles (EVs) increases, so does the volume and complexity of related data, such as vehicle type, model year, manufacturer, registration location, and ownership trends. Among the numerous data models that have been created over the years, the relational model—which is backed by well-known implementations such as Microsoft SQL Server, MySQL, and Oracle Databases —has predominated since the 1980s. These systems are collectively referred to as relational database management systems, or RDBMS. These systems are renowned for their robust querying capabilities, structured schema support, and strong consistency. But in recent years, relational databases' drawbacks have become more noticeable, particularly when working with dynamic, massive, or semi-structured data. Some of the primary challenges include limited horizontal scalability across dispersed environments, rigid schema structures, and inefficient modeling for highly interconnected datasets. Two worldwide trends have alerted the software and data engineering communities to these limitations. First, the centralization of data in vast distributed cloud systems run by companies like Amazon and Google amplifies the exponential growth in data volume, which is driven through sensors, systems, and users. Second, the increasing complexity and dependence of data —which has been accelerated by the growth of the Internet, Web 2.0 technologies, social networks, and the widespread availability of open APIs—causes traditional RDBMSs to occasionally fall short of the new requirements for flexible, scalable, and efficient data architectures. NoSQL databases, also referred to as non-relational databases are becoming more and more popular among enterprises that gather substantial volumes of semi-structured or unstructured data [32]. NoSQL systems have better horizontal scalability compared to commodity hardware and focus on processing huge datasets analytically [33]. Traditional SQL-based systems have reached their architectural boundaries as petabyte-scale data is processed by applications like social networking platforms, business intelligence, and big data analytics [34, 35, 36]. This limitation has driven the development of distributed and horizontally scalable NoSQL data stores as Google’s Bigtable [37], Apache

HBase, and Facebook's Cassandra [38]. Additionally, cassandra and Voldemort are examples of distributed key-value stores further demonstrate the efficiency and cost-effectiveness of NoSQL-based solutions [39]. One of the primary limitations of RDBMS is their difficulty in scaling efficiently with cloud apps, Web 2.0, grid computing, and data warehousing [40]. In contrast, NoSQL databases were designed to address these scalability concerns, offering schema flexibility, eventual consistency, and distributed storage and computation capabilities. As Pokorný (2011) points out, NoSQL platforms particularly excel in cloud computing contexts due to their horizontal scalability and high concurrency models [41]. Unlike RDBMS, however, most NoSQL systems do not guarantee full ACID compliance [42], trading strong consistency and transactions for availability and partition tolerance. The rise of Web 2.0 applications further motivated the shift toward NoSQL databases [43]. In such environments, data is often highly heterogeneous, comprising text, comments, images, videos, and source code, all of which challenge the rigid schema structure of relational databases. NoSQL databases offer agility and adaptability, allowing schema evolution without service interruption—a crucial feature for modern applications where adding or removing attributes must be quick and seamless [43]. These systems also support high-throughput indexing and massive concurrency, making them suitable for modern EV ecosystems and digital infrastructures that must process real-time, high-volume data streams [44]. Understanding the key components of both paradigms may be useful when choosing a storage engine for a particular solution, but as that solution develops over time, the needs may change and the current paradigm may not be able to meet them all. It would be advantageous to create a hybrid strategy in order to gain from the features of both paradigms because solutions that are restricted to only relational or non-relational paradigms are unable to utilize the advantages of the other one [45]. A hybrid database model, such as that found in the SQL and NoSQL paradigms, is an abstraction layer that sits on top of databases and integrates two or more different database models. Flexibility [46], improved performance, logical distribution, and web-friendly design [47] are some advantages of utilizing multiple database models in a system. In fact, some methods combine the NoSQL and SQL paradigms to create a hybrid database [48]. Nevertheless, there aren't many studies suggesting hybrid databases that integrate PostgreSQL and MongoDB. These databases are unique in their respective paradigms due to their characteristics. PostgreSQL was among the first databases to address geographical challenges. Because of its many spatial functions and high optimization for spatial queries, PostGIS, an extension of PostgreSQL, is highly relevant. In contrast, of the more than 225 NoSQL databases that are now on the market, MongoDB is the only NoSQL document-based database that allows point containment and line intersection searches. Additionally, Geoserver, an open source server for sharing geospatial data, is enabled to support both Database Management Systems (DBMS) since it incorporates a MongoDB data connection and publication component in version 2.11.4.

- **Machine Learning for Electric Vehicle Data**

Electric vehicles (EVs) are gaining momentum as a key pillar of sustainable transportation, driven by the global need to enhance urban air quality, decrease the use of fossil fuels and greenhouse gas emissions.As EV adoption increases, researchers and industry professionals are pursuing technological advancements aimed at enhancing performance and user experience. In particular, machine learning (ML) inclusion has emerged as a disruptive approach to optimize where EV systems manage energy use, battery health, vehicle performance, and efficiencies. Numerous dangers are captured by the vast and diverse statistics produced by EVs, including driving habits, battery load, location services, environment, and checking system diagnostics, are generally well captured and analyzed for machine learning models. ML for predictive modelling can lead to better forecasting of energy use, optimized charging, battery degradation prediction, and trip range reliability. For example, [49] illustrated the trend forecasting reliability of ensemble machine learning models for electric vehicle energy demand and use. Moreover, as pointed out by [50], some more sophisticated approaches, such as federated learning, can provide structured and private collaboration across distributed EV networks, advancing just-in-time decision making and resource allocation. The research highlighted in this review primarily examined ways that machine learning is transforming EV systems, including aspects of battery management systems, charging action analysis, communication efficiency and vehicle classification. Future possibilities for embedding ML-enabled solutions into complex, hybrid big data systems are discussed, along with challenges (for example, data privacy, effective model scaling and heterogeneous infrastructure). The application of machine learning to electric vehicle (EV) systems has been the subject of a substantial amount of recent study in particular and battery management and energy efficiency. In [51], for example, suggested how machine learning might help forecast battery degradation and allow real-time tracking of effective battery management based on various models for machine learning, for instance, SVM, RF and Neural Networks. The researchers in [52] and [53] also examined how Recurrent neural networks and Long Short-Term

Memory models could offer a more precise estimation of State of Charge (SoC) and State of Health batteries with improved accuracy over traditional electrochemical-based models. In [54] and [55], they examined Autoencoders and Isolation Forests in the context of predictive maintenance to identify anomalous performance in a battery and possible early tracking of failure. In [56], they conducted further studies focusing on the efficacy of ML in enhancing SoC and SoH estimation while providing less complexity. Furthermore, in [57] and [58] also studied the prediction of energy consumption with deep learning models such as LSTMs and demonstrated how this prediction can be enhanced when external data was included (weather and traffic) Reinforcement learning can be an applied to develop dynamic pricing parameters for reducing charging behaviour on peak load pressure, as evidenced by [59]. Moreover, [60] examined possible approach of decentralisation to improve privacy and scalability of charging infrastructure with federated learning. Collectively, these studies illustrate that machine learning is having an expanding influence in turn on the EV ecosystem and building more user-friendly, efficient, and intelligent vehicle practices.

2. METHODOLOGY

Developing a hybrid big data architecture that integrates SQL and NoSQL database types for storing and analyzing electric vehicle population datasets is the goal of this study. The architecture includes machine learning models to classify electric vehicles into their respective categories based on associated features: model year, make, vehicle class, and fuel type. The process has four major phases: data acquisition, machine learning model building, hybrid data management and storage, and performance assessment. This study focuses on achieving scalable efficient processing of the mixed data formats while also increasing the classification accuracy of electric vehicles and allowing for data-driven understanding of EV adoption trends.

2.1. Data Source

The source for data used in this project, was obtained from Data.gov, which is the official open government data platform for the United States. More specifically, the Department of Licensing in Washington State gave us the dataset called "Electric Vehicle Population Data", which gives us all the complete and updated information of electric vehicles (EVs) that are registered in the state of Washington. There are several formats available for the dataset:

- CSV format – used as the structured data source to simulate SQL-based storage and queries.
- JSON format – used to represent semi-structured NoSQL data for flexibility in storage and fast retrieval of dynamic information

2.2. Data Processing

Given that raw structured and semi-structured data was hybrid formatted for storage and analysis, the aim of the data processing stage was to clean and normalize it. Relevant features were first manually selected based on importance to the study's objectives. This stage aimed to increase the quality of the ensuing analysis by reducing the dataset's dimension and eliminating elements that were thought to be unnecessary or redundant.

The selected dataset underwent a series of preprocessing procedures. Duplicate records and records with missing values were flagged and removed to ensure trustworthiness and consistency. Categorical variables were changed to numerical representations to allow the data to be more compatible with analytical and machine learning processes. To discover the data's underlying structure, exploratory data analysis was used. Exploratory data analysis helped identify patterns, distributions, and relationships among important features. The exploratory stage aided the decision-making processes through statistical visualizations namely boxplots, distribution plots, and correlation heatmaps to gather insights. By the end of the preprocessing stage, the dataset was organized, cleaned, and ready for inclusion in hybrid data management systems, and the inception of intelligent modeling processes. The detailed processing of records improved data quality, eliminated noise from the records, and increased the potency of subsequent integration with machine learning algorithms.

ISSN: Printed: **3078-9656, Online: 3078-9664**, paper ID: 106

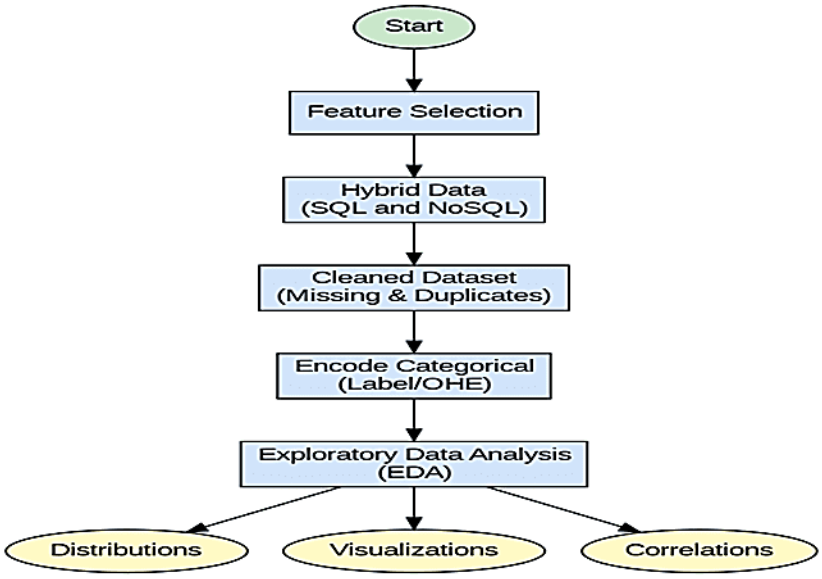

Figure2. Data Processing Using Hybrid Methods

An Exploratory Data Analysis (EDA) was conducted as part of the data processing step in order to have a deeper comprehension of the dataset's properties and structure. In this step, statistical illustrations were created, such as distribution plots for numerical features, boxplots to show the existence of outliers, and count plots for the categorical variables. A correlation heatmap was used to display the relationships between the features. Important background information for the feature selection and model design stages was supplied by the illustrations. Figures 3 through 4 present the findings of this investigation.

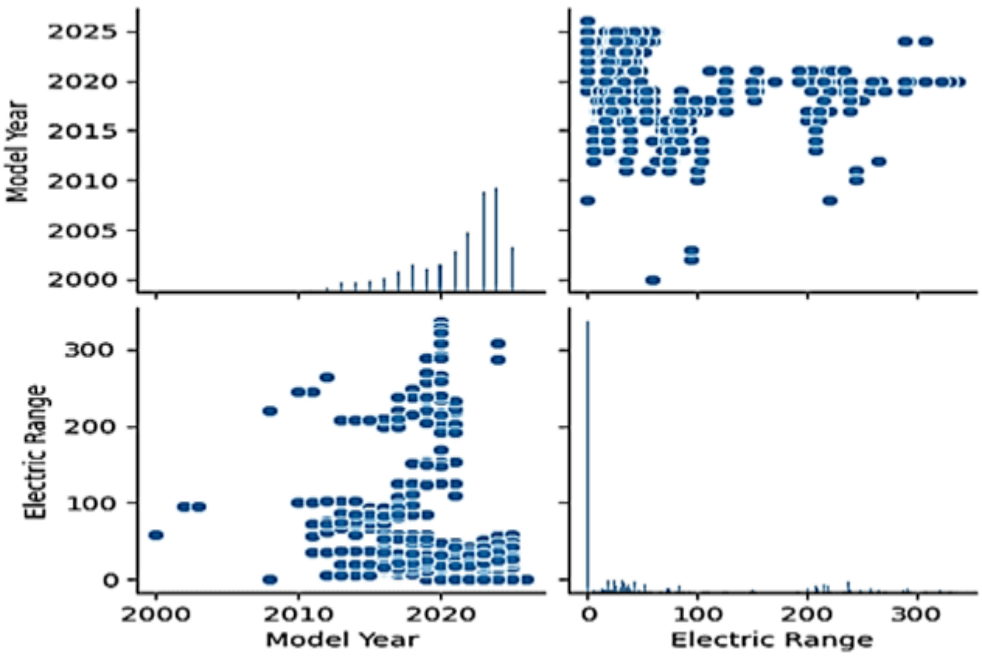

Figure 3. Distribution and Visualization of Features

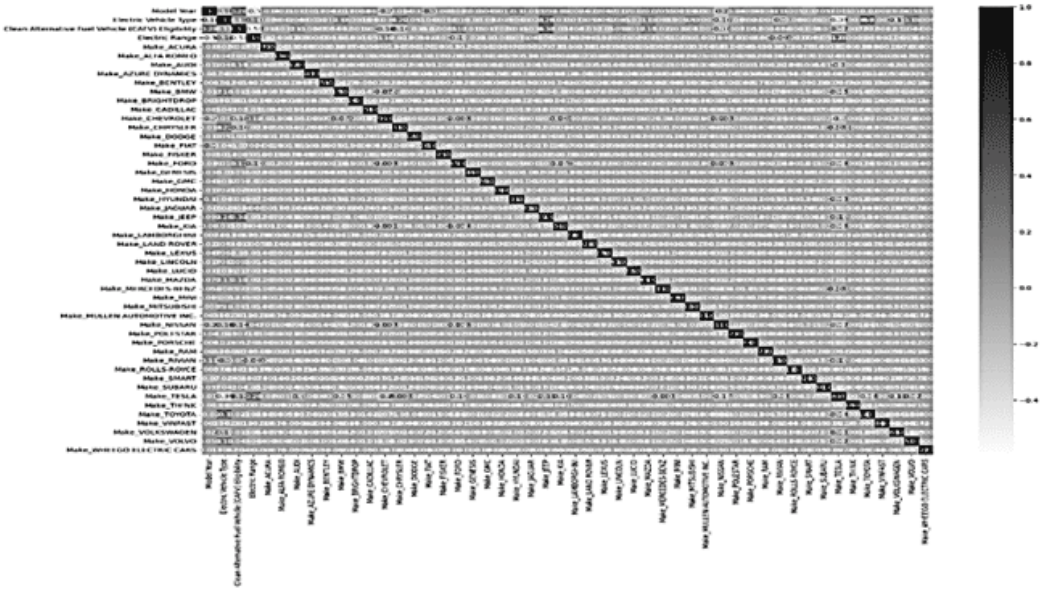

Figure 4. Feature Correlation Heatmap

After completing the procedures for feature selection, encoding, and data cleansing, the finished dataset was prepared for the integration of machine learning methods. After preprocessing, a total of 135,765 records were present. The data was split to training (70%) and testing (30%) sections in order to train and assess the model. This division maintains a distinct subset for objective assessment while guaranteeing that the amount of data used to train the model is sufficient. To ensure uniformity and reproducibility throughout the experiments, a fixed random seed was employed.

2.3 Machine learning algorithms

This study made use of a number of machine learning algorithms to classify different kinds of electric vehicles using semi-structured and structured data. The algorithms that were chosen for classification tasks—Decision Tree, K-Nearest Neighbours, Support Vector Machine, Random Forest, Logistic Regression, and AdaBoost—represent a variety of learning paradigms and have complementary advantages.

- Decision Tree

Coming upon a tree with leaves and branches, where the leaves represent characteristics for the finished product—is how the DT algorithm operates. The dataset is separated into subsets during the training phase so that the tree can be created. Because DT able to manage both category and numerical data. It can be applied to problems involving regression and classification. It is frequently used for a variety of applications because missing values in a dataset are known to be particularly useful. For identifying the output, the most precise and pure nodes are the decision nodes, are created during training by recursively branching nodes and this the decision tree's function. The input determines which branches will be used to identify the outcome or final forecast. A Decision Tree splits the input space based on feature thresholds. For example, suppose we have a tree that makes decisions based on two features: x1 and x2. The logic of the tree might be written as:

$$f(x) = \begin{cases} 1 & \text{if } x_1 \leq 2.5 \text{ and } x_2 > 3.0 \\ 0 & \text{if } x_1 \leq 2.5 \text{ and } x_2 \leq 3.0 \\ 1 & \text{if } x_1 > 2.5 \text{ and } x_2 \leq 4.5 \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

- K-Nearest Neighbors

Since KNN assumes that similar objects are close to one another, the output is dependent on how closely the input resembles the data used to train the model. Its ability to manage both numerical and categorical data makes It works well for both classification and regression. The modal class of the K points that are closest to the input is the output in classification. The regression's output is the mean of the values of the K points that are closest to the input Euclidean distance, the most popular KNN metric, determines the linear separation between two points in n-dimensional space.

$$d(\mathrm{x}, \mathrm{x}_i) = \sqrt{\sum_{j=1}^{n} \left(x_j - x_{ij}\right)^2} \quad (2)$$

- Support Vector Machine

As it learns from the training data, the Support Vector Machine (SVM) efficiently separates data points from several classes by creating a decision boundary, which can be a hyperplane in greater dimensions or a line in two dimensions. The algorithm aims to identify the optimal boundary by maximizing margin, this represents the Euclidean distance between each class's closest data points, or support vectors, and the separating hyperplane. SVM is an ideal margin classifier since it concentrates on optimizing this margin to improve ability of the classifier to apply to unidentified data. Once the boundary is established, the classification of new instances depends on which side of the hyperplane they fall. SVM uses the kernel trick, a mathematical technique that converts input data to allow for a higher-dimensional feature's linear separation space when there is no linear separability of the data. The resulting hyperplane is then mapped back to the original input space to serve as the decision boundary. The SVM decision function in the linear case is represented as:

$$\mathrm{f(x)} = \mathrm{w^T x} + \mathrm{b} \quad (3)$$

- Random Forest

DT serves as the foundation for the ML algorithm known as RF. Consequently, for DT, it can be applied to both regression and classification. Instead of finding a single DT, RF discovers several. As a result, RF generates several classification or regression with a tree structure. The final output for classification is the mode of the results from each individual tree, whereas the final output for regression is based on the mean of the output of all the singular trees. As a result, RF selects or averages the best class that the distinct trees predict.
Mathematically, the predicted class yˆ for an input x is given by:

$$\hat{y} = \arg\max_{c \in \mathcal{C}} \sum_{t=1}^{T} \mathbb{I}(h_t(\mathbf{x}) = c) \quad (4)$$

- Logistic Regression

For applications involving binary classification, one popular machine learning approach is logistic regression. Since the main objective is to estimate a binary outcome's probability based on a specified collection of input characteristics. The logistic function, which is another name for the sigmoid function, is used by the model to translate the attributes of the input into a probability value that ranges from 0 to 1. Through reducing the disparity between the anticipated and actual class labels probabilities, maximum likelihood estimation enhances the model parameters. Let y be a binary output variable and x be an input feature vector with n dimensions in mathematics. The model of logistic regression assumes that the conditional probability of y=1 given x is:

$$p(y = 1 \mid \mathrm{x}) = \frac{1}{1+\exp\left(-\mathrm{w}^T \mathrm{x} - b\right)} \quad (5)$$

- AdaBoost

AdaBoost is a very effective ensemble learning method for classification in the form of supervised learning as well as regression in the form of supervised learning. By combining weak learners, we can create a strong classifier which will give us a better chance of becoming increasingly accurate. AdaBoost iteratively reweighs the training samples and modifies the weight of those samples accordingly, which is intended to give more focus on those

original training samples that were misclassified, which then allows weak learners to be appropriately trained. The resulting model, with contributions to predictive performance in areas such as cybersecurity, computer vision, healthcare, and finance, is reduced to a weighted sum of the weak learners. The AdaBoost classifier will mathematically be expressed as:

$$H(\mathrm{x}) = \mathrm{sign}(\sum_{t=1}^{T} \alpha_t h_t(\mathrm{x})) \quad (6)$$

The system architecture for the analysis of electric vehicle data presented in this paper follows a hybrid big data processing approach architecture to account for a variety of structured and semi-structured data sources by merging the SQL and NoSQL formats. The system architecture starts with hybrid data collection and data preprocessing, which is followed by feature transformation and selection, then machine learning classification algorithm performance is assessed separately identifying different kinds of electric vehicles. Figure 5 presents the overall architecture and the workflow of the proposed system.

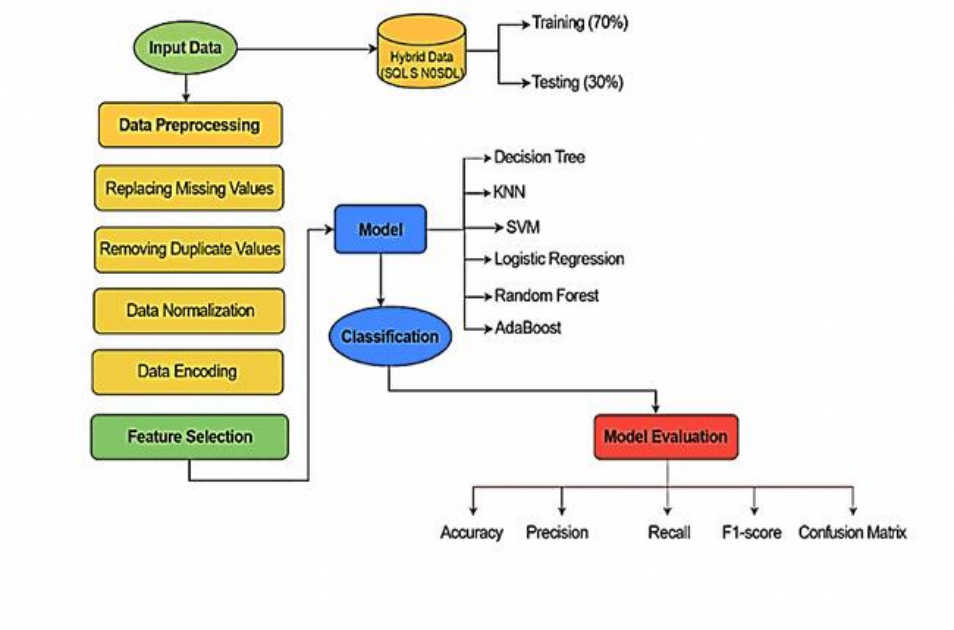

Figure 5. Hybrid System Architecture for Electric Vehicle Data Analysis

As part of the system architecture, a number of independent machine learning classification algorithms were trained and evaluated to assess how well they predicted the various types of electric vehicles based on certain features. The learned algorithms include AdaBoost, Random Forest, K-Nearest Neighbours, Support Vector Machine, Decision Tree, and Logistic Regression. The performance of each classifier was assessed using the same preprocessed dataset and common performance standards with the F1-score, confusion matrix, recall, accuracy, and precision.

The system architecture for the analysis of electric vehicle data presented in this paper follows a hybrid big data processing approach to handle a variety of structured and semi-structured data sources by merging the SQL and NoSQL formats. The architecture begins with hybrid data collection and preprocessing stages, including missing value handling, normalization, encoding, and feature selection. Following the preprocessing phase, machine learning classification algorithms are applied to predict different types of electric vehicles based on selected features. Figure 6 presents a simplified flow of the system components from big data sources through SQL/NoSQL

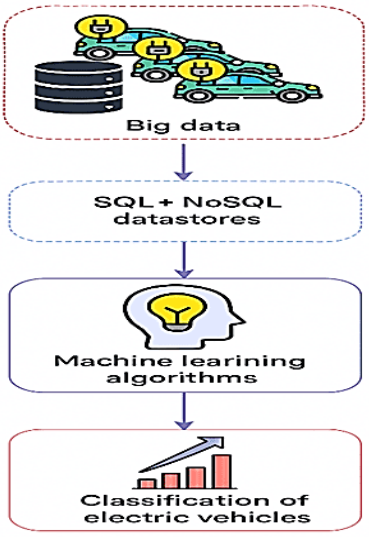

Figure 6. Architecture of the Proposed Electric Vehicle Classification System5

3. RESULTS AND DISCUSSION

A selection of algorithms for supervised machine learning was implemented to evaluate the efficacy of the proposed hybrid data architecture using data preprocessed to include features from both structured and semi-structured data sources. The models were independently trained and validated in a consistent manner, and consisted of several teaching techniques, AdaBoost, Random Forest, K-Nearest, Support Vector Machine, Decision Tree, and Logistic Regression. All algorithms were evaluated through important performance indicators, as accuracy, precision, recall, F1-score, and confusion matrix metrics. The proposed evaluation framework demonstrated each algorithm's ability to classify different types of electric vehicles in a heterogeneous data environment, as well as enabled a fair and comprehensive comparison of the models. The results are valuable in determining the best implementation for practical EV data analytics and each method's advantages and disadvantages.

- Accuracy

Accuracy is the frequency with which the model produces accurate forecasts, or the proportion of accurate forecasts of each and every prediction

.

$$\text{Accuracy} = (TP + TN)/(TP + FP + TN + FN) \quad (7)$$

- Precision

Classification accuracy by itself does not reveal which output classes are being accurately predicted when there are more than two. Precision is a better evaluation metric in these situations. Precision is the percentage of accurately anticipated positive outcomes among all projected positive outcomes.

$$\text{Precision} = TP / (TP + FP) \quad (8)$$

- Recall

Recall quantifies how well positive cases are predicted the percentage of true positives as determined by the model that were accurately predicted.

$$\text{Recall} = TP / (TP + FN) \quad (9)$$

- F1-Score

The F1-score, a statistic that integrates recall and accuracy, is especially important for models where false negatives as well as false positives can have detrimental effects. The model's performance can be balanced by calculating the precision and recall harmonic means.

$$\text{F1-Score} = \frac{2\times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (10)$$

The key performance metrics determined for each machine learning algorithm used to classify different types of electric vehicles are shown in Table 1. These metrics consist of F1-score, recall, accuracy, and precision. They highlight the trade-offs between various prediction error types and provide a comprehensive evaluation and comparison of each algorithm's performance in accurately predicting the vehicle classes.

Table 1. Evaluation of Classification Model Performance Comparison

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Decision Tree	99.99%	100%	100%	100%
KNN	99.99%	100%	100%	100%
SVM	91.22%	91%	91%	91%
Random Forest	99.99%	100%	100%	100%
Logistic Regression	91%	91%	91%	91%
AdaBoost	99.94%	100%	100%	100%

To better comprehend the prediction activities of each classification model, confusion matrices were created for each model after the performance comparison in Table 1. By displaying the accuracy and inaccuracy of each algorithm's classification of various electric vehicle kinds, these matrices offer a graphic assessment of false positives, false negatives, real positives, and true negatives. While the off-diagonal elements show misclassifications, the diagonal elements show accurate predictions. Below are the confusion matrices for every model:

Plotting the curve because it enabled us to evaluate the Receiver Operating Characteristic (ROC) the algorithms' performance beyond simple accuracy metrics. As shown in Figure 7, to demonstrate each model's level of performance can distinguish between classes, the ROC curve shows the True Positive Rate (sensitivity) in proportion to the rate of false positives across a variety of classification thresholds.

Among the tested models, Decision Tree, KNN, Random Forest, and AdaBoost all achieved very high accuracy rates (99.94%–99.99%) achieving high accuracy scores, recall, and F1-score. However, the top-performing model was determined to be Random Forest due to its high stability, strong generalization ability, and robustness to imbalanced or noisy data. Additionally, it provides feature importance insights, making it the most suitable choice for deployment within the practical system prototype for classifying electric vehicle types.

Figure 8 presents the Precision-Recall (PR) curves for all classifiers. The results indicate that the Decision Tree, KNN, Random Forest, and AdaBoost classifiers achieved near-perfect performance, each with an average precision of 1.00, demonstrating excellent classification capability. In contrast, the SVM and Logistic Regression models yielded lower average precision scores of 0.86 and 0.89, respectively, indicating a more moderate trade-off between precision and recall.

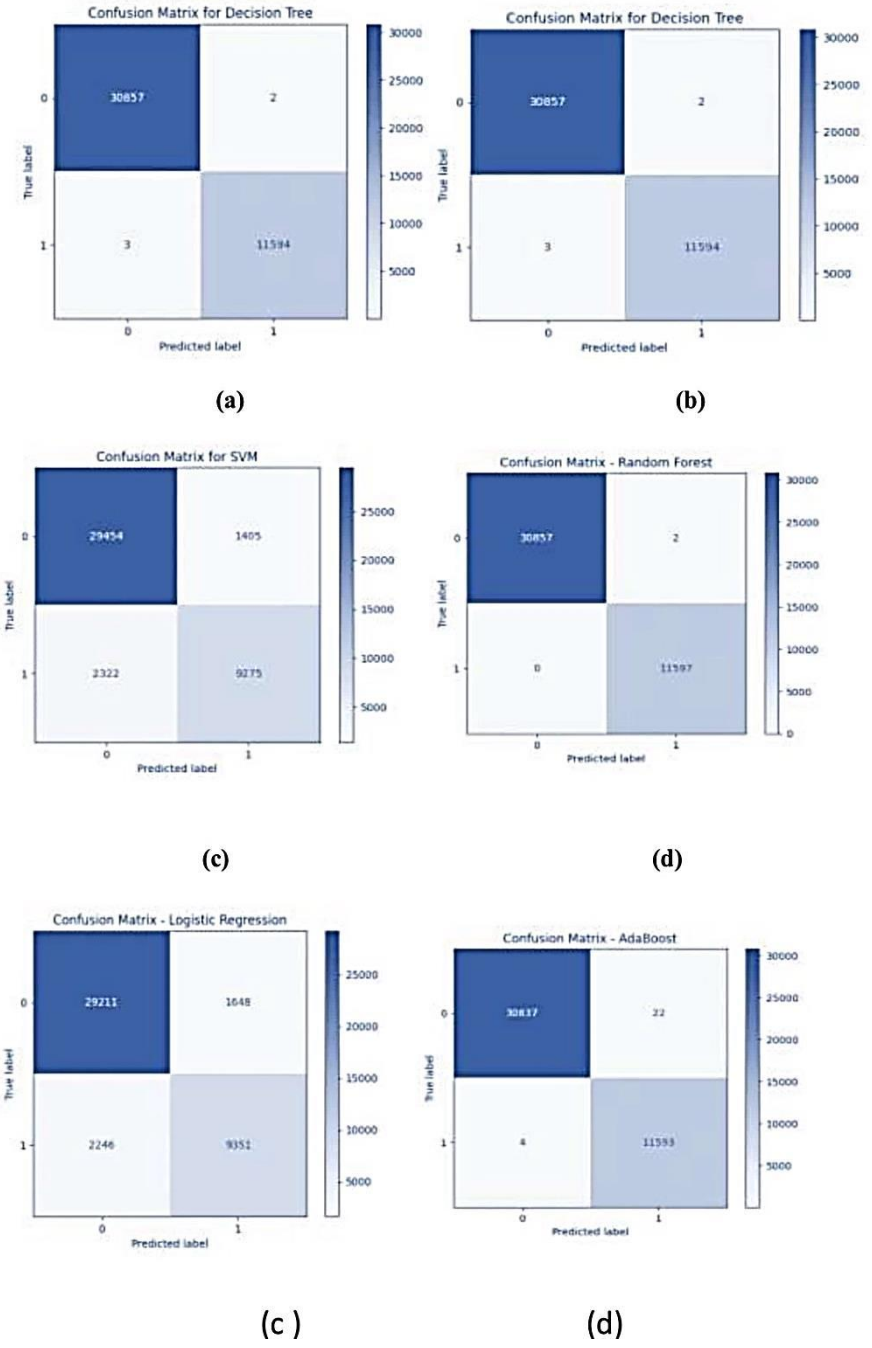

Figure 6. Confusion matrices of (a) Decision Tree, (b) KNN, (c) SVM, (d) Random Forest, (e) Logistic Regression, (f) AdaBoost

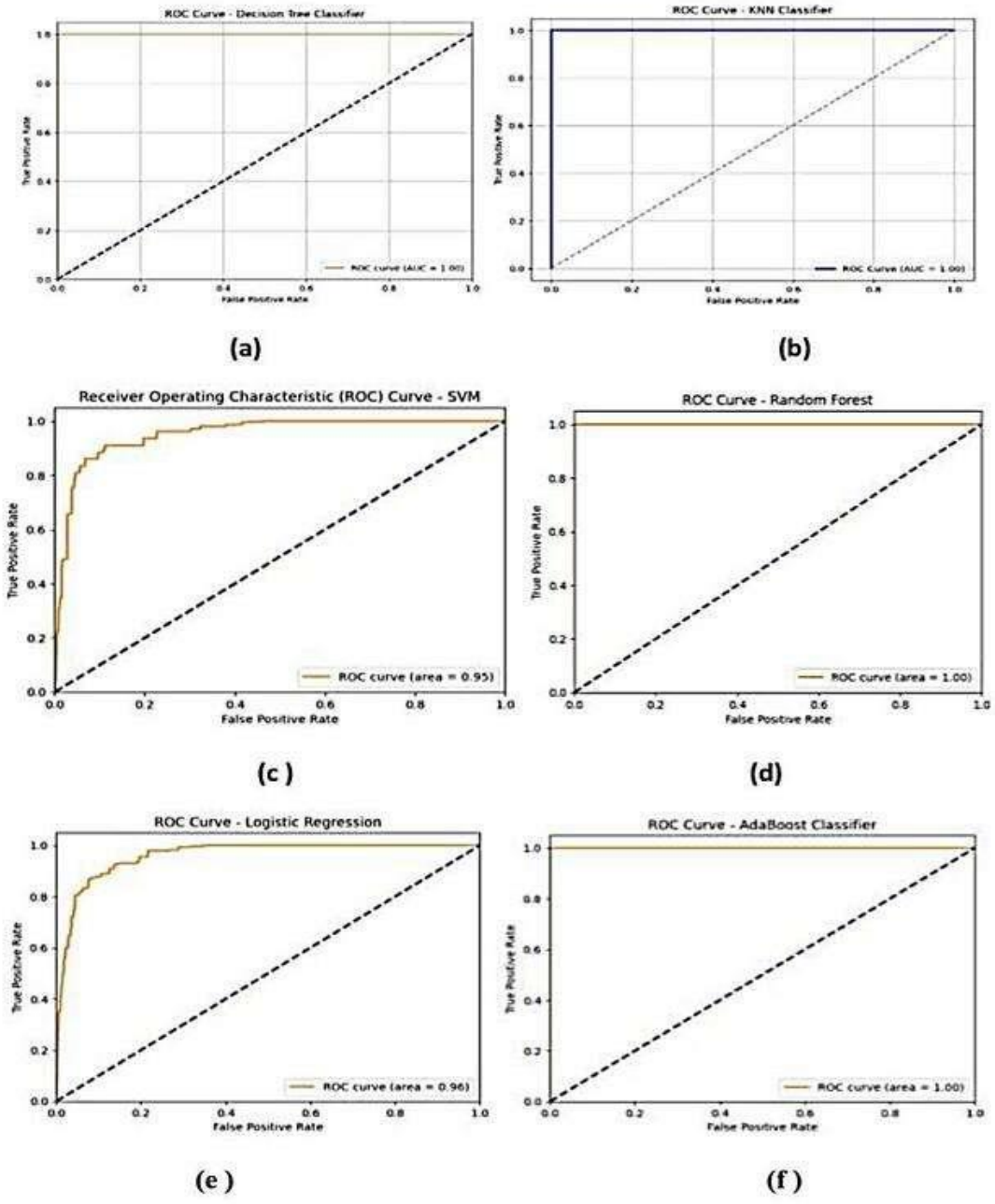

Figure 7. ROC curves of (a) Decision Tree, (b) KNN, (c) SVM, (d) Random Forest, (e) Logistic Regression, (f) AdaBoost

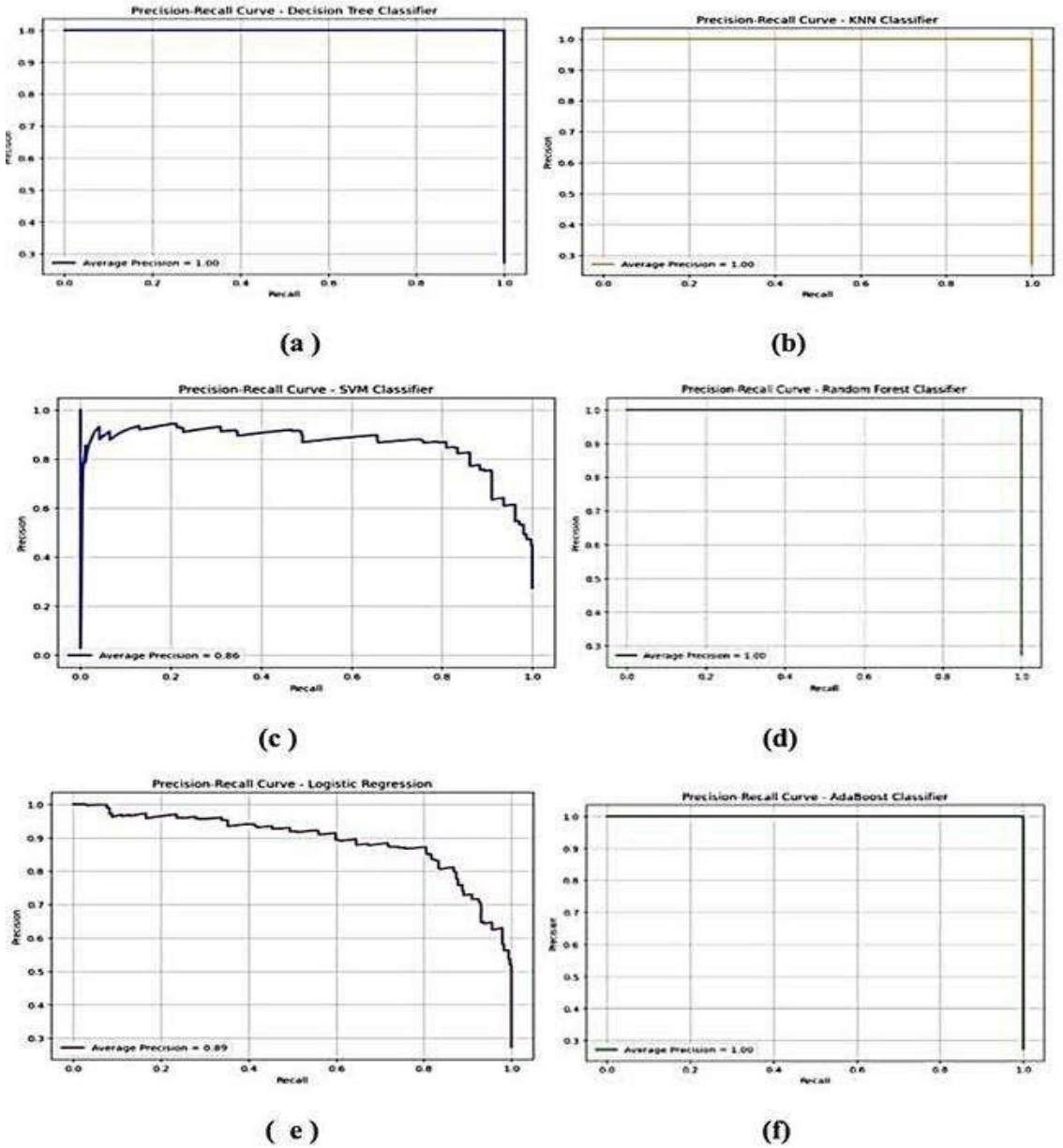

Figure 8. Precision-Recall Curve (a) Decision Tree, (b) KNN, (c) SVM, (d) Random Forest, (e) Logistic Regression, (f) AdaBoost

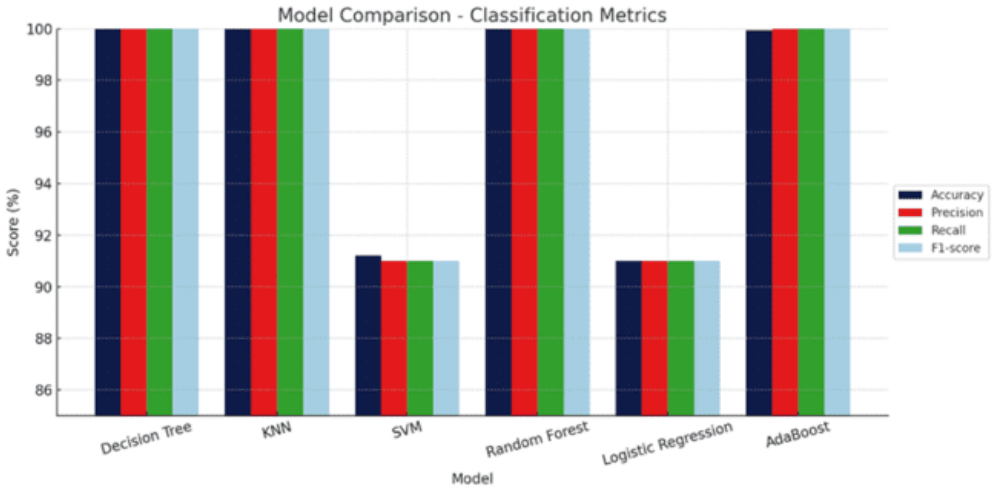

Figure 9. Model Comparison

As shown in Figure 9, classification metrics are compared across models, highlighting the superior performance of Decision Tree, KNN, Random Forest, and AdaBoost, while SVM and Logistic Regression showed comparatively lower scores.

Figure 10. Electric Vehicle Type Using Random Forest Classifier

3. Conclusion

A comprehensive Hybrid Big Data Architecture capable of ingesting multiple datasets was achieved by successfully merging relational (SQL) and non-relational (NoSQL) databases to work together. Enabled by this merged architecture, data could now be combined, stored, and processed together to form one environment for data analysis and knowledge extraction. The research utilized data mining techniques to explore and characterize patterns and structure within the dataset. In this study, a number of machine learning models were employed for classification, as the Random Forest, AdaBoost (KNN) classifier, Decision Tree, (SVM), and Logistic Regression

model. Each machine learning model was evaluated using F1 score, recall, accuracy, and precision. All classifiers performed remarkably well with Decision Tree, KNN and Random Forest classifiers, obtaining accuracy of 99.99% and precision, recall and F1 score of 100%. AdaBoost was close behind with 100% accuracy and 99.94% precision, recall and F1-score. SVM and Logistic Regression gave low scores of about 91% for each evaluating metric. Of all the various models tested, it was discovered that Random Forest was the highest-performing model overall. Its quality of extrapolation, high stability, and resistance to noisy/unbalanced data were critical to its success. Additionally, Random Forest provides feature importance information, which helps to interpret the model, as well as an understanding of the underlying relationships represented in the data. To sum up, this research affirms the usefulness of machine learning to facilitate registered decisions and the effectiveness of hybrid big data systems in processing multiple datasets. The architecture proposed provides a flexible and scalable model for a comprehensive range of other use cases needing real-time data integration and prediction beyond the classification of electric vehicles. A promising possibility for future research is expanding the hybrid data ecosystem with different types of data, especially since this study successfully developed a hybrid big data architecture that utilized both SQL and NoSQL data management models. This ecosystem approach could potentially include unstructured and semi-structured data, including additional text records, sensor logs (such as from charging stations), vehicle telematics, and even multimedia data assets such as images and videos. By integrating potentially useful hybrid data type into the existing system, a more holistic architecture can evolve to support new intelligent applications. Beyond providing greater flexibility and scalability to the system, this improvement would also position it to integrate with various domains including, connected infrastructure, mobility services, and smart cities.
next (based on result and discussion).

REFERENCES

[1] S. Hosseini and M. D. Sarder, "Development of a Bayesian network model for optimal site selection of electric vehicle charging station," International Journal of Electrical Power & Energy Systems, vol. 105, pp. 110–122, 2019. doi: 10.1016/j.ijepes.2018.08.011
[2] N. Kannan and D. Vakeesan, "Solar energy for future world: A review," Renewable and Sustainable Energy Reviews, vol. 62, pp. 1092–1105, 2016. doi: 10.1016/j.rser.2016.05.022
[3] Y. M. Wang, S. F. He, D. G. Zamora, X. H. Pan, and L. Martinez, "A large scale group three-way decision-based consensus model for site selection of electric vehicle charging stations," Expert Systems with Applications, vol. 214, p. 118884, 2023. doi: 10.1016/j.eswa.2022.119107
[4] L. Zhang and Q. D. Qin, "China's electric vehicle policies: Evolution, comparison, recommendation," Transportation Research Part A: Policy and Practice, vol. 110, pp. 57–72, 2018. doi: 10.1016/j.tra.2018.02.012
[5] J. Linn and V. McConnell, "Interactions between federal and state policies for reducing vehicle emissions," Energy Policy, vol. 126, pp. 507–517, 2019. https://doi.org/10.1016/j.enpol.2018.11.045
[6] V. Gass, J. Schmidt, and E. Schmid, "Analysis of alternative policy instruments to promote electric vehicles in Austria," Renewable Energy, vol. 61, pp. 96–101, 2014. https://doi.org/10.1016/j.renene.2013.06.013
[7] S. F. Tie and C. W. Tan, "A review of energy sources and energy management system in electric vehicles," Renewable and Sustainable Energy Reviews, vol. 20, pp. 82–102, 2013. https://doi.org/10.1016/j.rser.2012.11.024
[8] D. Vienneau et al., "Years of life lost and morbidity cases attributable to transportation noise and air pollution: a comparative health risk assessment for Switzerland in 2010," International Journal of Hygiene and Environmental Health, vol. 218, no. 1, pp. 1–10, 2015. doi: 10.1016/j.ijheh.2014.08.004.
[9] K. Kieckhäfer, K. Wachter, and T. S. Spengler, "Analyzing manufacturers' impact on green products' market diffusion – the case of electric vehicles," Journal of Cleaner Production, vol. 162, pp. 89–105, 2017. https://doi.org/10.1016/j.jclepro.2017.06.114
[10] K. S. Gallagher and E. Muehlegger, "Giving green to get green? Incentives and consumer adoption of hybrid vehicle technology," Journal of Environmental Economics and Management, vol. 61, no. 1, pp. 1–15, 2011. https://doi.org/10.1016/j.jeem.2010.10.003
[11] S. Wang, J. Wang, J. Li, et al., "Policy implications for promoting the adoption of electric vehicles: Do consumers' knowledge, perceived risk, and financial incentive policy matter?," Transportation Research Part A: Policy and Practice, vol. 117, pp. 58–73, 2018. https://doi.org/10.1016/j.tra.2018.08.014

[12] C. Chellaswamy, A. Jayaprakash, et al., "Future renewable energy option for recharging full electric vehicles," Renewable and Sustainable Energy Reviews, vol. 76, pp. 824–838, 2017. https://doi.org/10.1016/j.rser.2017.03.118
[13] J. McLaren, L. Gagnon, et al., "CO_2 emissions associated with electric vehicle charging: The impact of electricity generation mix, charging infrastructure availability, and vehicle type," The Electricity Journal, vol. 29, no. 5, pp. 72–83, 2016. https://doi.org/10.1016/j.tej.2016.06.005
[14] E. A. Nanaki and C. J. Koroneos, "Comparative economic and environmental analysis of conventional, hybrid and electric vehicles – The case study of Greece," Journal of Cleaner Production, vol. 53, pp. 261–266, 2013. https://doi.org/10.1016/j.jclepro.2013.04.010
[15] J. Meckling, N. Kelsey, E. Biber, and J. Zysman, "The politics of technology bans: Industrial policy competition and green goals for the auto industry," Energy Policy, vol. 132, pp. 803–815, 2019. https://doi.org/10.1016/j.enpol.2019.06.020
[16] J. Linn, E. Muehlegger, and D. Rapson, "Interactions between federal and state policies for reducing vehicle emissions," Energy Policy, vol. 129, pp. 1170–1178, 2019. https://doi.org/10.1016/j.enpol.2019.03.003 https://doi.org/10.1016/j.enpol.2019.03.003
[17] L. V. White, K. Tong, and M. Wachs, "Why are charging stations associated with electric vehicle adoption? Untangling effects in three United States metropolitan areas," Energy Research & Social Science, vol. 86, p. 102423, 2022. https://doi.org/10.1016/j.erss.2021.102423
[18] L. Maybury, J. McLean, and P. Thornley, "Mathematical modelling of electric vehicle adoption: A systematic literature review," Transportation Research Part D: Transport and Environment, vol. 109, p. 103342, 2022. https://doi.org/10.1016/j.trd.2022.103342
[19] M. Coffman, P. Bernstein, and S. Wee, "Electric vehicles revisited: A review of factors that affect adoption," Transportation Reviews, vol. 37, no. 1, pp. 79–93, 2017. https://doi.org/10.1080/01441647.2016.1194721
[20] R. G. Newell and D. Raimi, "US federal government subsidies for clean energy: Design choices and implications," Energy Economics, vol. 78, pp. 106–122, 2019. https://doi.org/10.1016/j.eneco.2018.11.010
[21] S. He, Y. Wang, J. Xu, and L. Martínez, "A large-scale group decision-making model for optimal site selection of electric vehicle charging stations," Expert Systems with Applications, vol. 213, p. 118884, 2023. https://doi.org/10.1016/j.eswa.2022.119107
[22] M. Masmoudi, H. B. Ghezala, and M. Bouzeghoub, "Big data analytics in smart transportation: A systematic mapping study," Journal of Big Data, vol. 8, no. 1, pp. 1–26, 2021. https://doi.org/10.1186/s40537-021-00460-4
[23] F. Bonomi, R. Milito, J. Zhu, and S. Addepalli, "Fog computing and its role in the Internet of Things," in Proceedings of the First Edition of the MCC Workshop on Mobile Cloud Computing, 2012, pp. 13–16. https://doi.org/10.1145/2342509.2342513
[24] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, "Internet of Things for smart cities," IEEE Internet of Things Journal, vol. 1, no. 1, pp. 22–32, 2014. https://doi.org/10.1109/JIOT.2014.2306328
[25] H. Ning and H. Liu, "Cyber-physical-social-thinking space based science and technology framework for the Internet of Things," Science China Information Sciences, vol. 58, no. 3, p. 1–19, 2015. https://doi.org/10.1007/s11432-014-5209-2
[26] M. A. Khan and K. Salah, "IoT security: Review, blockchain solutions, and open challenges," Future Generation Computer Systems, vol. 82, pp. 395–411, 2018. https://doi.org/10.1016/j.future.2017.11.022
[27] X. Xu, C. He, and P. Xu, "Big data analytics in transportation: A review," Journal of Transportation Systems Engineering and Information Technology, vol. 18, no. 3, pp. 29–42, 2018. https://doi.org/10.1016/j.jttse.2018.02.004
[28] M. F. Hasan, T. A. Rashid, and A. S. Al-Fedaghi, "A review on Internet of Things (IoT): Security and privacy requirements and the solution approaches," Journal of Network and Computer Applications, vol. 181, p. 103001, 2021. https://doi.org/10.1016/j.jnca.2021.103001
[29] M. Wolf, "Cyber-physical systems security for the Internet of Things," Proceedings of the IEEE, vol. 104, no. 5, pp. 1066–1081, 2016. https://doi.org/10.1109/JPROC.2016.2534698
[30] S. Abhishek and R. Rathi, "Big data analytics in transportation systems: A survey," International Journal of Transportation Science and Technology, vol. 7, no. 4, pp. 295–310, 2018. https://doi.org/10.1016/j.ijtst.2018.11.001

[31] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A survey," Computer Networks, vol. 54, no. 15, pp. 2787–2805, 2010. https://doi.org/10.1016/j.comnet.2010.05.010?utm_source=chatgpt.com
[32] H. Boyes, B. Hallaq, J. Cunningham, and T. Watson, "The industrial Internet of Things (IIoT): An analysis framework," Computers in Industry, vol. 101, pp. 1–12, 2018. https://doi.org/10.1016/j.compind.2018.04.015?utm_source=chatgpt.com
[33] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A vision, architectural elements, and future directions," Future Generation Computer Systems, vol. 29, no. 7, pp. 1645–1660, 2013. https://doi.org/10.1016/j.future.2013.01.010?utm_source=chatgpt.com
[34] K. Ashton, "That 'Internet of Things' thing," RFID Journal, vol. 22, no. 7, pp. 97–114, 2009.
[35] J. Manyika et al., "Big data: The next frontier for innovation, competition, and productivity," McKinsey Global Institute, pp. 1–137, 2011.
[36] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, "Context aware computing for the Internet of Things: A survey," IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 414–454, 2014. https://doi.org/10.1109/SURV.2013.042313.00197?utm_source=chatgpt.com
[37] D. Evans, "The Internet of Things: How the next evolution of the Internet is changing everything," Cisco Internet Business Solutions Group (IBSG), pp. 1–11, 2011.
[38] A. Botta, W. De Donato, V. Persico, and A. Pescapé, "Integration of cloud computing and Internet of Things: A survey," Future Generation Computer Systems, vol. 56, pp. 684–700, 2016. https://doi.org/10.1016/j.future.2015.09.021?utm_source=chatgpt.com
[39] R. Khan, S. U. Khan, R. Zaheer, and S. Khan, "Future Internet: The Internet of Things architecture, possible applications and key challenges," in Proceedings of the 10th International Conference on Frontiers of Information Technology (FIT), IEEE, 2012, pp. 257–260. https://doi.org/10.1109/FIT.2012.53?utm_source=chatgpt.com
[40] J. Holler et al., From Machine-to-Machine to the Internet of Things: Introduction to a New Age of Intelligence, Academic Press, 2014.
[41] C. M. Medaglia and A. Serbanati, "An overview of privacy and security issues in the Internet of Things," in The Internet of Things, Springer, 2010, pp. 389–395. https://doi.org/10.1007/978-3-642-19157-2_22?utm_source=chatgpt.com
[42] P. P. Ray, "A survey of IoT cloud platforms," Future Computing and Informatics Journal, vol. 1, no. 1-2, pp. 35–46, 2016. https://doi.org/10.1016/j.fcij.2016.07.001?utm_source=chatgpt.com
[43] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A survey on enabling technologies, protocols, and applications," IEEE Communications Surveys & Tutorials, vol. 17, no. 4, pp. 2347–2376, 2015. https://doi.org/10.1109/COMST.2015.2444095?utm_source=chatgpt.com
[44] R. Want, B. N. Schilit, and S. Jenson, "Enabling the Internet of Things," Computer, vol. 48, no. 1, pp. 28–35, 2015. https://doi.org/10.1109/MC.2015.19?utm_source=chatgpt.com
[45] M. Chiang and T. Zhang, "Fog and IoT: An overview of research opportunities," IEEE Internet of Things Journal, vol. 3, no. 6, pp. 854–864, 2016. https://doi.org/10.1109/JIOT.2016.2579198?utm_source=chatgpt.com
[46] [46] F. Wortmann and K. Flüchter, "Internet of Things," Business & Information Systems Engineering, vol. 57, no. 3, pp. 221–224, 2015. https://doi.org/10.1007/s12599-015-0383-3?utm_source=chatgpt.com
[47] A. Whitmore, A. Agarwal, and L. Da Xu, "The Internet of Things—A survey of topics and trends," Information Systems Frontiers, vol. 17, no. 2, pp. 261–274, 2015. https://doi.org/10.1007/s10796-014-9489-2?utm_source=chatgpt.com
[48] S. Li, L. Da Xu, and S. Zhao, "The Internet of Things: A survey," Information Systems Frontiers, vol. 17, no. 2, pp. 243–259, 2015 https://doi.org/10.1007/s10796-014-9492-6?utm_source=chatgpt.com .
[49] L. Da Xu, W. He, and S. Li, "Internet of Things in industries: A survey," IEEE Transactions on Industrial Informatics, vol. 10, no. 4, pp. 2233–2243, 2014. https://doi.org/10.1109/TII.2014.2300753?utm_source=chatgpt.com
[50] A. Zanella et al., "Internet of Things for smart cities," IEEE Internet of Things Journal, vol. 1, no. 1, pp. 22–32, 2014. https://doi.org/10.1109/MIC.2010.20?utm_source=chatgpt.com
[51] G. Kortuem, F. Kawsar, V. Sundramoorthy, and D. Fitton, "Smart objects as building blocks for the Internet of Things," IEEE Internet Computing, vol. 14, no. 1, pp. 44–51, 2010. https://doi.org/10.1016/j.jmsy.2018.01.006?utm_source=chatgpt.com

[52] F. Tao, Q. Qi, A. Liu, and A. Kusiak, "Data-driven smart manufacturing," Journal of Manufacturing Systems, vol. 48, pp. 157–169, 2018 https://doi.org/10.1109/MS.2016.18?utm_source=chatgpt.com .
[53] M. Weyrich and C. Ebert, "Reference architectures for the Internet of Things," IEEE Software, vol. 33, no. 1, pp. 112–116, 2016. https://doi.org/10.1109/MS.2016.18?utm_source=chatgpt.com
[54] R. Minerva, A. Biru, and D. Rotondi, "Towards a definition of the Internet of Things (IoT)," IEEE Internet Initiative, vol. 1, pp. 1–86, 2015.
[55] A. S. Shrouf, J. Ordieres-Meré, A. García-Sánchez, and M. Ortega-Mier, "Optimizing the production scheduling of a single machine to minimize total energy consumption costs," Journal of Cleaner Production, vol. 67, pp. 197–207, 2014. https://doi.org/10.1016/j.jclepro.2013.12.043?utm_source=chatgpt.com
[56] H. Sundmaeker, P. Guillemin, P. Friess, and S. Woelfflé, Vision and Challenges for Realising the Internet of Things, Cluster of European Research Projects on the Internet of Things, European Commission, 2010.
[57] O. Vermesan and P. Friess, Internet of Things: Converging Technologies for Smart Environments and Integrated Ecosystems, River Publishers, 2013.
[58] P. P. Ray, "A survey on Internet of Things architectures," Journal of King Saud University-Computer and Information Sciences, vol. 30, no. 3, pp. 291–319, 2018. https://doi.org/10.1016/j.jksuci.2016.10.003?utm_source=chatgpt.com
[59] J. Lin, W. Yu, N. Zhang, X. Yang, H. Zhang, and W. Zhao, "A survey on Internet of Things: Architecture, enabling technologies, security and privacy, and applications," IEEE Internet of Things Journal, vol. 4, no. 5, pp. 1125–1142, 2017. https://doi.org/10.1109/JIOT.2017.2683200?utm_source=chatgpt.com
[60] Y. Sun, H. Song, A. J. Jara, and R. Bie, "Internet of Things and big data analytics for smart and connected communities," IEEE Access, vol. 4, pp. 766–773, 2016. https://doi.org/10.1109/ACCESS.2016.2568851?utm_source=chatgpt.com