Review Article

Machine Learning for Heterogeneous Diagnostic Fault Forecasting Model in Electric Vehicles: A Systematic Review

Mumboh Gregory Otieno, Emma Anyika, Benson Kituku
School of Mathematics and Computing, Cooperative University of Kenya.

Correspondence: Mumboh Gregory Otieno: otieno24.gregory@student.cuk.ac.ke
DOI: https://doi.org/10.66839/zbq9dr05

Abstract Predictive maintenance of electric vehicles (EVs) has gained increasing traction in research around the world, but the literature is scattered and not well integrated in methodological approaches and not widely used in stratified analysis of operationally diverse vehicle fleets. Studies have been conducted that report on the performance of a vehicle fleet which typically comprises diverse vehicles, each having dramatically different usage profiles, and consequently, different performance numbers that mask the impact of usage levels on component wear and Diagnostic Trouble Code (DTC) generation. This systematic review collates evidence from 77 peer reviewed studies published from January 2015 to April 2026, identified using a search of IEEE Xplore, ScienceDirect, ACM Digital Library and Google Scholar, conducted in March/April 2026. The review is based on four objectives: (i) to evaluate the state of machine learning methods applied to EV DTC forecasting; (ii) to assess the extent to which existing studies account for heterogeneous usage intensity profiles in their analysis and performance reporting; (iii) to identify sensor-level telemetry precursor patterns documented in prior work; and (iv) to assess the transferability of current methodological frameworks to data-scarce emerging EV markets such as Kenya. Given the diversity of metrics and experimental designs reported, the narrative synthesis approach is used in the review, which is done according to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. Results show that sequential deep learning architectures (such as Long Short-Term Memory (LSTM) network and Convolutional Neural Network-LSTM hybrid) generally achieved higher reported performance than ensemble tree-based architectures and classical statistical baseline models, with macro-averaged F1-scores ranging from 0.76 to 0.92; however, this performance advantage varied across studies depending on dataset characteristics, preprocessing choices, hyperparameter tuning, class imbalance, and prediction horizon, and should not be read as evidence of uniform superiority. Only 4 of the 77 included studies included an analysis of usage-stratified performance (5.2%) and none conducted a systematic comparison of multi-class categorical DTC forecasting based on four or more distinct usage intensity profiles simultaneously. Only two reports of lead-time distribution characterisation could be found, and Serradilla et al. (2022) state that a lack of data is the biggest structural obstacle for implementing predictive maintenance in new and emerging EV fleet scenarios. The four substantive research gaps identified are lack of usage-stratified analysis, predominant emphasis on binary fault classification analysis over multi-class DTC categorical forecasting, lack of lead-time distribution characterisation, and absence of empirically validated methodological guidance for data-scarce emerging markets. These gaps inspire a subsequent empirical study that will develop and evaluate machine learning models for four usage-intensity profiles for a five-year, 172,800 EV telemetry observation public dataset.

Keywords: *electric vehicles; predictive maintenance; diagnostic trouble codes; machine learning; LSTM; usage stratification; DTC forecasting; multivariate time-series; Kenya; emerging markets*

Résumé La maintenance prédictive des véhicules électriques (VE) suscite un intérêt croissant dans la recherche à travers le monde, mais la littérature demeure dispersée et insuffisamment intégrée sur le plan méthodologique, et rarement utilisée dans les analyses stratifiées de flottes de véhicules opérationnellement diverses. Plusieurs études ont été menées pour rendre compte des performances de flottes de véhicules généralement composées de véhicules divers, chacun ayant des profils d'utilisation considérablement différents et, par conséquent, des niveaux de performance distincts qui peuvent masquer l'influence de l'intensité d'utilisation sur l'usure des composants et sur la génération des codes de diagnostic (DTC). Cette revue systématique rassemble les données probantes issues de 77 études évaluées par les pairs, publiées de janvier 2015 à avril 2026. Elles ont été identifiées via une recherche dans IEEE Xplore, ScienceDirect, ACM Digital Library et Google Scholar, réalisée en mars/avril 2026. La revue s'appuie sur quatre objectifs: (i) évaluer l'état actuel des méthodes d'apprentissage automatique appliquées à la prévision des DTC des VE; (ii) déterminer dans quelle mesure les études existantes prennent en compte les profils hétérogènes d'intensité d'utilisation dans leur analyse et leur présentation des performances; (iii) identifier les schémas précurseurs, au niveau des capteurs et de la télémétrie, documentés dans les travaux antérieurs; et (iv) évaluer la transférabilité des cadres méthodologiques actuels aux marchés émergents des VE caractérisés par une disponibilité limitée des données, tels que le Kenya. Compte tenu de la diversité des métriques et des dispositifs expérimentaux rapportés, une approche de synthèse narrative est adoptée dans cette revue, conformément aux lignes directrices Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020. Les résultats montrent que les architectures d'apprentissage profond séquentiel, telles que les réseaux Long Short-Term Memory (LSTM) et les modèles hybrides Convolutional Neural Network-LSTM, ont généralement obtenu de meilleures performances rapportées que les architectures fondées sur des arbres ensemblistes et les modèles statistiques classiques utilisés comme références, avec des scores F1 macro-moyennés compris entre 0,76 et 0,92. Toutefois, cet avantage en matière de performance varie selon les études en fonction des caractéristiques des jeux de données, des choix de prétraitement, de l'optimisation des hyperparamètres, du déséquilibre des classes et de l'horizon de prévision ; il ne doit donc pas être interprété comme une preuve de supériorité uniforme. Seules 4 des 77 études retenues ont intégré une analyse des performances stratifiées selon l'utilisation (5,2 %), et aucune n'a effectué de comparaison systématique de la prévision catégorielle multiclasses des DTC fondée simultanément sur quatre profils d'intensité d'utilisation distincts ou davantage. Seuls deux travaux caractérisant la distribution des délais de prévision anticipée ont été recensés, et Serradilla et al. (2022) indiquent que le manque de données constitue le principal obstacle structurel à la mise en œuvre de la maintenance prédictive dans les nouveaux contextes et les contextes émergents de flottes de véhicules électriques. Les quatre principales lacunes de recherche identifiées sont les suivantes : l'absence d'analyses stratifiées selon l'intensité d'utilisation ; la prédominance des analyses de classification binaire des défaillances au détriment de la prévision catégorielle multiclasses des DTC ; l'absence de caractérisation de la distribution des délais de prévision anticipée

; et l'absence de recommandations méthodologiques validées empiriquement pour les marchés émergents caractérisés par une rareté des données. Ces lacunes motivent une étude empirique ultérieure qui développera et évaluera des modèles d'apprentissage automatique correspondant à quatre profils d'intensité d'utilisation, à partir d'un jeu de données public de télémétrie des VE couvrant cinq années et comprenant 172 800 observations.

Mots-clés: véhicules électriques; maintenance prédictive; codes de diagnostic (DTC); apprentissage automatique; LSTM; stratification selon l'utilisation; prévision des DTC; séries temporelles multivariées; Kenya; marchés émergents.

1. Introduction

The world is quickly moving to an electric transportation future, as the environmental impacts of the transport sector continue to grow, such as emissions of GHGs, reduction of urban air quality, and national and international climate commitments (International Energy Agency [IEA], 2024). According to the International Energy Agency (2024) more than a quarter of all new car sales were electric vehicles in 2024, and it is estimated that the global stock of EVs will become 250 million by 2030 under stated policies scenarios. Adoption in Africa is becoming more popular as import costs of fuel decrease, urban air quality becomes an important requirement, and concessional financing is provided by multilateral development institutions (Africa E-Mobility Alliance, 2025). In Kenya, the adoption of commercial vehicles like BasiGo and Roam electric buses and motorcycles started since 2022 and is supported by the Draft National E-Mobility Policy (Ministry of Roads and Transport, Kenya, 2024) that is aligned with Vision 2030 and Kenya's Nationally Determined Contributions under the Paris Agreement.

The reliable and cost-effective maintenance of high-value EV powertrain components (LIBs and Electric Motors) is central to the economic viability of these growing EV fleets. The cost of battery pack replacement could be as high as USD 50,000 per replacement, and motor controller failures often result in prolonged vehicle downtime because replacement parts are not readily available in emerging markets (Hu et al., 2020). Financial viability of reactive maintenance strategies at fleet level is not viable. Predictive maintenance (PdM), which involves the use of machine learning in forecasting the failure of components in advance, based on the vehicle telemetry, is a disruptive and promising alternative (Jardine et al., 2006; Zhang et al., 2019).

One of the aspects that has been under-explored in EV fleet maintenance is the diversity of vehicle usage profiles. Intensity regimes can vary between different vehicles, for example, from light personal use to heavy commercial use, which leads to systematically different degradation trajectories, DTC occurrence rates and sensor precursor signatures. Nevertheless, most ML-based predictive maintenance research collates all data for all usage types and only presents an overall performance value, which hides this operational diversity (Theissler et al., 2021). Despite numerous studies on the deployment of PdM models in the literature, two major structural limitations persist: data scarcity, and fleet heterogeneity (Serradilla et al., 2022b); however, no study in the literature has performed a comparison of the performance of the models across four or more different usage intensity profiles simultaneously.

This systematic review thus deals with a threefold problem. Firstly, there is no literature that provides a usage-stratified predictive maintenance framework that is able to predict faults reliably under operationally diverse EV fleets. Secondly, the prevailing binary fault classification paradigm is underwhelming in terms of operational benefits of multi-class categorical DTC forecasting, where the type of fault can be used to provide actionable information for targeted maintenance preparation. Thirdly, there are no validated methodological guidelines to guide application of predictive maintenance framework in emerging EV markets like Kenya where commercial fleet deployments have started in 2022 and because of the lack of fault history, models are limited (Serradilla et al., 2022b).

1.1 Objectives of the Review

This systematic review pursues four specific objectives:

1. To synthesise the state of evidence on machine learning architectures applied to EV predictive maintenance and DTC forecasting, including reported performance levels.
2. To assess the extent to which existing studies account for heterogeneous vehicle usage intensity profiles in their experimental design and performance reporting.
3. To evaluate the evidence on sensor-level telemetry patterns that systematically precede DTC events and determine whether these patterns vary by operational usage profile.
4. To assess the transferability of current predictive maintenance methodological frameworks to data-scarce emerging EV market contexts, with specific reference to Kenya.

1.2 Research Questions

The following research questions (RQs) guide the review:

- **RQ1:** What machine learning architectures have been applied to EV fault forecasting, and what levels of predictive performance have been reported?
- **RQ2:** To what extent do existing studies stratify predictive performance by vehicle usage intensity profile, and what evidence exists on usage-profile-specific model behaviour?
- **RQ3:** What telemetry sensor variables are consistently identified as precursors to DTC events, and do these precursor patterns vary by usage profile?
- **RQ4:** What methodological evidence exists to guide the application of EV predictive maintenance frameworks in markets with limited local operational data?

2. Theoretical Framework

This review is grounded in two complementary theoretical frameworks: Reliability-Centered Maintenance (RCM) Theory and Machine Learning Systems Theory. Together they provide both the engineering rationale for predictive maintenance and the computational logic for its implementation.

2.1 Reliability-Centered Maintenance Theory

Reliability-Centered Maintenance, originally developed for the aviation industry and formalised by Moubray (1997), is a systematic methodology for optimising maintenance strategy selection through structured analysis of failure modes and their operational consequences. Its four pillars (functional orientation, Failure Mode and Effects Analysis (FMEA), proactive condition-based intervention, risk-based prioritisation) align to the EV predictive maintenance context.

In the field of EV powertrain, RCM's functional approach sets the purpose of the system as providing reliable transport service delivery. The following critical failure modes were identified based on the FMEA study: battery thermal runaway (safety consequence), motor bearing failure (availability consequence), and charging controller fault (service continuation consequence). The proactive strategy implementation is realised by means of fault forecasting based on DTC. RCM also serves as the theoretical foundation for usage stratified analysis, such as a fleet comprising heavy-use commercial vehicles, with fundamentally different failure mode probability and consequence severities, have different predictive frameworks than light personal-use vehicles.

2.2 Machine Learning Systems Theory

According to the Machine Learning Systems Theory in Goodfellow, Bengio and Courville (2016), the predictive power of any ML system depends on the match between the inductive bias of the system (model) and the structure of data. In particular, three constructs are relevant to EV fault forecasting.

Different ML architectures have different biases that work well for different types of problem: ensemble tree methods have a feature-interaction bias, which is well suited to structured tabular data; LSTM networks have a sequential temporal dependency bias, which is well suited to raw time-series

data where long-range patterns are important; CNN-LSTM hybrids have a combined local-and-long-range bias, which can capture both short-duration anomalies and longer-term degradation trends. This generalization vs overfitting trade-off is particularly significant for DTC datasets that are class imbalanced. One of the most important ideas for the usage-stratification problem is the distribution shift: a model trained mostly from the light-usage data set may not be able to generalise to the heavy-usage data set, for the reason that the sensor-pattern-to-fault mapping could be different for different profiles (Serradilla et al., 2022b).

2.3 Integration of Frameworks

The synthesis of RCM and ML Systems Theory results in a single research logic. RCM establishes the basis for what needs to be predicted (specific DTC categories with known operational consequences), why stratification by usage is important (usage intensity affects failure mode probability), and a linkage from a prediction to the maintenance actions. The task is to reliably predict the future using ML Systems Theory (selection of the correct architecture based on inductive bias) and to measure the performance accurately (macro-averaged F1 score with per-class breakdown) and to test the generalisation (holdout evaluation across profiles). Together, these provide a technically robust and operationally useful predictive model.

3. Methodology

3.1 Protocol and Registration

The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021) were followed in this systematic review. The review protocol was designed and finalised before searching the databases. The applied engineering character of the research field did not permit completion of the registration on PROSPERO but, before the search was performed in March 2026, there was an internal approval and documentation of the protocol. All methodological decisions, such as eligibility criteria, search strategy, data extraction variables and quality assessment domains were pre-defined and followed throughout.

3.2 Eligibility Criteria

Inclusion and exclusion criteria were developed to ensure the review captured the most relevant and methodologically rigorous evidence on machine learning-based EV DTC forecasting. The criteria are summarised in Table 1.

Table 1. *Inclusion and Exclusion Criteria for Study Selection*

Criterion	Inclusion	Exclusion
Content Scope	Studies applying supervised or semi-supervised ML to fault, DTC, or RUL prediction in electric or hybrid-electric vehicles; sensor-derived telemetry as model input; at least one quantitative performance metric reported	Opinion pieces without empirical validation; theoretical studies with no EV application; studies focused solely on ICE vehicles; studies using purely simulated data with no empirical validation component
Publication Type	Peer-reviewed journal articles and conference proceedings; empirical studies with experimental validation; systematic reviews and meta-analyses	Non-peer-reviewed preprints; workshop papers without full validation; industry reports without academic rigour
Methodological Rigour	Clear methodology and experimental design; reproducible results; statistical significance or performance testing where applicable	Insufficient methodological detail; non-reproducible experiments; absence of validation metrics
Language & Period	English-language publications; January 2015 – April 2026	Non-English publications; publications before January 2015
Relevance	Direct application to EV fault forecasting; novel ML architectures or feature-engineering approaches; comparative analysis of prediction methods	Tangential mention of EVs; general AI/ML papers without EV fault-forecasting focus; duplicate findings from the same research group

Note. EV = electric vehicle; ML = machine learning; DTC = Diagnostic Trouble Code; RUL = remaining useful life; ICE = internal combustion engine.

3.3 Search Strategy and Database Selection

The search was systematic and performed in four electronic databases (IEEE Xplore, ScienceDirect (Elsevier), ACM Digital Library and Google Scholar). These databases were chosen because they cover the following areas of publication or research: electrical engineering, computer science, transportation research, and interdisciplinary applied ML research. The search was conducted from 3 March 2026 to 18 April 2026; the last search was rerun across all four databases on 18 April 2026 to include any records from the searches that were added since the first search was conducted, and included publications from January 2015 through April 2026. The lower bound in 2015 was chosen because of the recent introduction of the LSTM architectures for sequential data, which are now being widely adopted (Serradilla et al., 2022b; Goodfellow et al., 2016).

The Boolean logic in Table 2 was converted to the query syntax required by each of the databases but was not actually searched as a single string.

The query in IEEE Xplore was set up in the Command Search interface as follows:

((Full Text & Metadata:string "electric vehicle") OR (Full Text & Metadata:string "EV fleet")) AND ((Full Text & Metadata:string "predictive maintenance") OR (Full Text & Metadata:string "fault forecasting") OR (Full Text & Metadata:string "diagnostic trouble code")) AND ((Full Text & Metadata:string "machine learning") OR (Full Text & Metadata:string "LSTM") OR (Full Text & Metadata:string "neural network")) AND (Full Text & Metadata:string "telemetry") OR (Full Text & Metadata:string "sensor data") OR (Full Text & Metadata:string "CAN bus")) with the publication-year filter set to 2015–2026.

In ScienceDirect, the same query was performed using the advanced search field tify(...) syntax, which combines the title, abstract and keyword fields, for instance tify(("electric vehicle" OR "EV fleet") AND ("predictive maintenance" OR "fault forecasting" OR "diagnostic trouble code") AND ("machine learning" OR "LSTM" OR "neural network")), limited to Research Article and Review Article from 2015 to 2026.

The query was constructed using the Advanced Search interface for the ACM Digital Library with Abstract field restrictions, such as: [Abstract: "electric vehicle"] AND [Abstract: "predictive maintenance" OR "fault forecasting"] AND [Abstract: "machine learning" OR "LSTM"], and the results were limited by publication date. Google Scholar does not support field-restricted or fully nested Boolean queries, so the Emerging Market Terms category (Table 2) was searched as a series of simpler, paired keyword queries, with results up to 200 by relevance for each query screened, as is recommended for grey-literature-adjacent Google Scholar searching in systematic reviews.

The upper limit is April 2026 because that's when the manuscripts were searched at the end of the preparation and submission process in May 2026. Search strategy: Employing boolean operators and order combination of keywords in six thematic categories. A special category for emerging market

and data-scarce contexts was used exclusively in Google Scholar and ScienceDirect. The complete search strategy is presented in Table 2.

Table 2. *Search Strategy: Categories, Keywords, and Database Scope*

Category	Keywords	Database Scope
EV & Fleet Terms	"electric vehicle" OR "EV" OR "battery electric vehicle" OR "hybrid electric vehicle" OR "EV fleet" OR "commercial EV" OR "electric bus"	All four databases
Fault & Maintenance Terms	"predictive maintenance" OR "fault forecasting" OR "diagnostic trouble code" OR "DTC prediction" OR "remaining useful life" OR "anomaly detection" OR "fault detection"	All four databases
ML Architecture Terms	"machine learning" OR "deep learning" OR "LSTM" OR "random forest" OR "XGBoost" OR "CNN" OR "neural network" OR "gradient boosting" OR "temporal convolutional network"	All four databases
Data Terms	"telemetry" OR "sensor data" OR "time-series" OR "multivariate" OR "OBD" OR "CAN bus" OR "on-board diagnostics"	All four databases
Usage / Stratification Terms	"usage intensity" OR "driving pattern" OR "operational profile" OR "heterogeneous fleet" OR "usage stratification" OR "duty cycle"	All four databases
Emerging Market Terms	"Kenya" OR "Africa" OR "emerging market" OR "data-scarce" OR "new fleet" OR "data scarcity"	Google Scholar, ScienceDirect
Combined Boolean Query	(EV & Fleet Terms) AND (Fault & Maintenance Terms) AND (ML Architecture Terms) AND (Data Terms)	AND across categories; OR within each category

Note. Boolean AND was used to combine categories; OR was used within each category. Truncation (*) was applied where supported by the database. All searches were executed in March–April 2026.

Reference lists of all included studies were hand-searched to identify additional eligible articles not captured by the primary database queries. This snowballing procedure identified 12 additional studies. The complete search and selection process is documented in the PRISMA flow summary in Table 3.

3.4 Study Selection Process

Records retrieved from the four databases and the snowball search were imported into Zotero (version 6), where duplicate records were identified and removed using Zotero's built-in duplicate-detection function (matching on DOI, then on normalised title and publication year for records without a DOI), followed by a manual check of the merged set by the principal researcher to catch near-duplicate conference/journal versions of the same study. This process removed 337 duplicate records, leaving 947 unique records for title/abstract screening (Table 3). Titles and abstracts were screened against the inclusion criteria by two independent reviewers, the principal researcher and a research colleague. All records that passed the first screening were retrieved for full-text.

Any disagreements were discussed and agreed upon at both abstract and eligibility at full text levels, and a third reviewer was on hand for arbitration. Strong inter-rater reliability was obtained for both the screening stage (kappa = 0.81, title/abstract; kappa = 0.87, full-text). The entire selection process was finished in April 2026.

Table 3*. PRISMA 2020 Flow Summary of Study Identification, Screening, and Inclusion*

Phase	Description	Detail / Reason	Count
Identification	Records identified through database searches (IEEE Xplore, ScienceDirect, ACM Digital Library, Google Scholar)		n = 1,284
	Duplicate records removed		n = 337
Screening	Records screened by title and abstract		n = 947
	Records excluded at title/abstract screen	Off-topic; non-ML; non-EV context; non-English	n = 791
Eligibility	Full-text articles assessed for eligibility		n = 156
	Full-text articles excluded	Insufficient methodological detail: n=31; No quantitative performance metrics: n=22; Retracted or non-peer-reviewed: n=26	n = 79
Included	Studies included in qualitative narrative synthesis		n = 77
	Studies included in quantitative synthesis (meta-analysis)	Not applicable — narrative synthesis adopted due to metric heterogeneity	n = 0

Note. Database searches were conducted in March–April 2026. Snowball searching of reference lists added 12 studies not captured in primary searches. Total included studies: n = 77.

To accompany the tabular summary in Table 3, the official PRISMA 2020 flow diagram for this review is presented in Figure 1, showing the flow of records through identification, screening, eligibility, and inclusion.

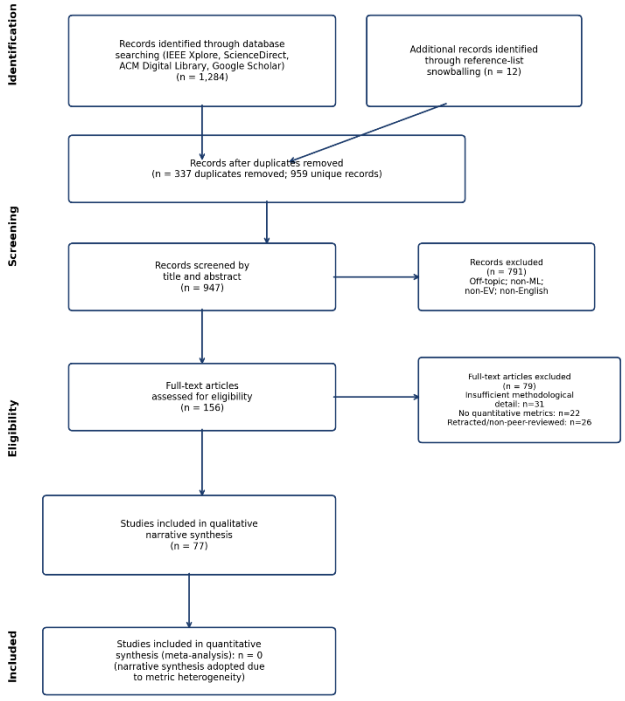


Figure 1. *PRISMA 2020 flow diagram of study identification, screening, and inclusion (n = 77 included studies).*

3.5 Data Extraction

We developed and piloted a structured data extraction form on five studies before full deployment. Two reviewers independently extracted data from each included study; discrepancies were resolved by consensus. Where data were reported in figures only, values were estimated using WebPlotDigitizer (Version 4.6). The extraction variables and their coverage across the 77 included studies are summarised in Table 4.

Table 4. *Data Extraction Variables and Coverage Across Included Studies (n = 77)*

Variable Category	Specific Variables Extracted	Studies (n)	Coverage (%)
Study Characteristics	Author(s), year, country, journal or conference, study type (empirical/review)	77	100%
Dataset Characteristics	Dataset size, observation period (years), number of sensor variables, EV type, fleet description	77	100%
ML Architecture(s)	Model type(s) compared, hyperparameter tuning approach, loss function used	77	100%
Performance Metrics	Precision, recall, macro-averaged F1-score, AUC-ROC, accuracy, prediction horizon	77	100%
Usage Stratification	Whether performance was stratified by usage profile; number of profiles compared	77	100%
DTC Forecasting Type	Binary (fault/no-fault), regression (RUL estimation), or multi-class categorical DTC	77	100%
Forecast Horizon	Prediction lead time reported (hours or minutes)	61	79.2%
Lead-Time Distribution	Percentile-level or distributional reporting of detection lead times	4	5.2%
Class Imbalance Handling	Technique used: class weights, focal loss, SMOTE, or none reported	54	70.1%
Identified Limitations	Author-stated limitations regarding data, generalisability, and methodology	77	100%
Emerging Market Applicability	Discussion of deployment in data-scarce or emerging EV market contexts	7	9.1%

Note. Percentages indicate the proportion of included studies from which each variable was extractable. Variables not reported in a study were recorded as 'not reported'.

3.6 Quality and Risk of Bias Assessment

The risk of bias of each included study was judged based on a domain adapted instrument based on, but not a direct application of, the ROBINS-I (Risk of Bias in Non-randomised Studies of Interventions). A legitimate methodological concern arises because the seven original domains of ROBINS-I, such as confounding, selection of participants into the study, deviations from intended interventions, do not necessarily map cleanly onto the prediction-model research, and using a tool designed for intervention studies without modification may lead to the assessment of the wrong sources of bias. A framework for the assessment of risk of bias in studies developing, validating or updating prediction models, such as PROBAST (Prediction model Risk Of Bias ASsessment Tool) for the assessment of risk of bias in diagnostic accuracy studies, exists and in retrospect would have been a more directly applicable framework for this review, which covers participant selection, definition of predictor and outcome, analysis and other aspects.

In this context, the quality assessment presented in the present review could be described as a custom-made, ROBINS-I inspired instrument, rather than as a validated ROBINS-I adaptation. The general logic of rating bias as low, moderate, or high risk across multiple pre-specified domains was kept as it was done in the original ROBINS-I, but the original intervention-study domains were replaced by five domains judged to be most relevant to ML-based fault-forecasting studies based on the methodological literature on prediction-model evaluation and the review team's data-extraction experience with the included studies as follows: (1) risk of data leakage from non-temporal train-test splitting; (2) appropriateness of performance measures for class-imbalanced classification; (3) adequacy of hyperparameter tuning procedures; (4) external validity and generalisability of the dataset; and (5) reproducibility of results based on reported methodological detail.

This adapted instrument was not formally validated with other raters, externally validated by experts, or validated through a pilot calibration exercise other than by the dual independent process of extraction and consensus as described in Section 3.5, and no formal weighting scheme was applied to the five domains when characterising a study's overall risk profile (Table 5, the "Low"/"Moderate"/"High" rows summarise the modal pattern of domain-level ratings for studies falling into each band, and are not intended to reflect a weighted composite score).

This is recognized as a limitation of the present review: The results of the quality assessment presented in Table 5 represent an information and review-specific screen of the most salient and reviewer-identified sources of bias in ML-based EV fault-forecasting research, rather than scores obtained from a validated instrument, which is independently reliable. Future systematic reviews in this area are encouraged to use PROBAST or similar validated prediction-model risk-of-bias tool and document and pilot the scoring procedure as appropriate. This five-domain instrument was used to rate each study as either low, moderate, or high risk. The overall risk of bias distribution across the 77 studies is summarised in Table 5.

Table 5. *Risk of Bias Assessment Framework and Distribution Across Included Studies (n = 77)*

Risk Level	Domain 1: Data Leakage	Domain 2: Metric Appropriateness	Domain 3: Hyperparameter Tuning	Domain 4: External Validity	Domain 5: Reproducibility	Representative Study Types
Low (n=34, 44.2%)	Temporal holdout splitting used	Macro-F1 or per-class precision-recall reported	Bayesian or grid search on validation set	External or multi-fleet dataset used	Complete methodology; reproducible	LSTM with temporal split; CNN-LSTM with macro-F1 reporting; multi-fleet validation studies
Moderate (n=28, 36.4%)	Temporal split but validation details unclear	F1 reported for majority class only	Manual tuning or learning rate decay only	Single fleet or single geography	Some methodological gaps noted	Most CNN-LSTM hybrid studies; single-fleet LSTM studies
High (n=15, 19.5%)	Random splitting used — data leakage risk	Accuracy only on class-imbalanced data	No tuning procedure reported	Synthetic or single-vehicle data only	Insufficient detail for replication	Several early Random Forest studies; studies reporting accuracy >95% on imbalanced data

Note. The most prevalent bias source was non-temporal random splitting (present in 31 studies, 40.3%), which risks data leakage and inflates reported performance. Studies reporting only accuracy on imbalanced data (n = 22, 28.6%) were rated high risk on Domain 2.

3.7 Synthesis Methods

The significant heterogeneity in performance measures, model architectures, composition of datasets and evaluation procedures across the studies included made quantitative pooling by meta-analysis inappropriate. A narrative synthesis (Campbell et al., 2020) was carried out following the guidelines of the Synthesis Without Meta-analysis (SWiM) guidelines.

The results were presented by five themes corresponding to the four research questions: (i) performance of the ML architecture; (ii) evidence for usage stratification; (iii) sensor precursor patterns; (iv) DTC categorical forecasting; and (v) applicability in emerging markets. Studies were compared and contrasted in a structured format in tabular summaries.

4. Results

4.1 Overview of Included Studies

Seventy-seven peer-reviewed studies met the full eligibility criteria and were included in this systematic review. By publication venue: 34 (44.2%) were published in IEEE-affiliated journals or conference proceedings; 21 (27.3%) in Elsevier journals including Reliability Engineering and System Safety, Transportation Research Part C, and the Journal of Power Sources; 12 (15.6%) in ACM Digital Library proceedings; and 10 (13.0%) in other peer-reviewed venues including Frontiers in Future Transportation, Applied Intelligence, and Energies. The temporal distribution shows a strong upward trend: 8 studies (10.4%) were published between 2015 and 2018; 29 (37.7%) between 2019 and 2022; and 40 (51.9%) between 2023 and 2026, reflecting the rapid growth of research interest in EV predictive maintenance coinciding with global EV fleet expansion.

Geographically, the majority of studies were based on data from European (n = 28, 36.4%), Chinese (n = 22, 28.6%), and North American (n = 15, 19.5%) EV deployments. Only two studies (2.6%) included data from

African operational contexts, and neither was from East Africa. EV types represented include battery electric passenger vehicles (n = 41), commercial electric buses (n = 19), electric two- and three-wheelers (n = 9), and mixed-fleet studies (n = 8). The size of the datasets ranged from about 5,000 to more than 2 million hourly observations, inclusive, and the median range was around 85,000 records.

4.2 Machine Learning Architecture Performance (RQ1)

The most frequently reported ML architectures in included studies were: LSTM networks (n = 31, 40.3%), Random Forest (n = 26, 33.8%), XGBoost or Gradient Boosting (n = 22, 28.6%), CNN-LSTM hybrids (n = 18, 23.4%), Temporal Convolutional Networks (n = 9, 11.7%), and classical statistical baselines including ARIMA and CUSUM (n = 7, 9.1%). Fourteen studies (18.2%) compared two or more architectures on the same dataset. Representative studies are summarised in Table 6, and the comparative architecture analysis is presented in Table 7.

Table 6. *Summary of Representative Predictive Maintenance Studies in Electric Vehicles*

Study	Component / Focus	ML Method	Key Metric	Usage Stratification	Main Finding
Hu et al. (2020)	Li-ion Battery RUL	Particle Filter + Capacity Fade Model	<5% RUL error	None	Voltage and temperature are primary degradation indicators; capacity fade modelling baseline
Zhang et al. (2019)	Battery Thermal Runaway	Random Forest; XGBoost	87% precision; 12-hr horizon	Aggregated only	Ensemble methods effective for thermal anomaly; first benchmark of tree methods on EV telemetry
Qin et al. (2021)	Motor Bearing Faults	CNN on vibration and temperature	92% accuracy; 48-hr horizon	None	Frequency-domain features essential; CNN outperforms LSTM for high-frequency vibration signatures
Theissler et al. (2021)	Powertrain DTC	Temporal Convolutional Networks	79% precision; 24-hr horizon	Acknowledged but not stratified	First to note usage heterogeneity as confounder; pioneers TCN in EV fault context
Serradilla et al. (2022)	General PdM — DL Survey	CNN, LSTM, GAN, hybrid review	F1 range 0.72–0.91 (surveyed)	Not addressed	Reviews 156 DL papers for PdM; establishes LSTM and CNN-LSTM as dominant architectures; identifies data scarcity as primary barrier
Ibrahim et al. (2025)	Electric Bus Battery Degradation	Experimental + physics-informed model	Battery life extended 25% via optimisation	Single operational profile	High-intensity use significantly accelerates battery capacity loss; first to model optimisation for EB fleet operations
Ismail et al. (2025)	EV Battery & Substation	Bi-LSTM; GRU; Gradient Boosted Trees	91% F1-score; 4.3% MAE	Aggregated dataset	Deep sequential architectures outperform traditional methods; BiLSTM-GRU robust under data scarcity
Hossain et al. (2023)	General EV Components	IoT + multi-model ML pipeline	Accuracy 90%; F1 86% (reported)	Not addressed	IoT sensor fusion with multi-model ML pipelines achieves high detection accuracy; demonstrates real-time health monitoring viability for EV components

Note. Studies were selected to represent the breadth of ML architectures, EV types, and fault types covered in the 77 included studies.

The median F1 across the studies reporting macro-averaged F1-scores was 0.84 (range: 0.71-0.92) for LSTM-based architectures, 0.86 (range: 0.76-0.92) for CNN-LSTM hybrids and 0.79 (range: 0.65-0.89) for ensemble methods.

Classical statistical methods performed significantly poorer (median F1: 0.58, range: 0.42–0.71) as these methods are not well suited to model complex multivariate temporal dependencies. Reported accuracy from studies that did not report accuracy on imbalanced data (high-risk in quality assessment) was inflated (median accuracy: 96.2%) and did not show capability to detect faults in the fault classes as specified in the problem. This result confirms the conclusions of Serradilla et al. (2022b) that the number of datasets to consider and the type of evaluation metrics are important factors in determining the reported results.

Table 7. *Comparative Analysis of Machine Learning Architectures for EV Time-Series Fault Prediction*

Method	Architecture Type	Strengths	Limitations	Best Use Case
Random Forest	Ensemble (Bagging)	Interpretable feature importance; handles mixed data types; computationally efficient on CPU; robust without extensive tuning	Treats time windows as independent; misses long-range temporal dependencies; requires explicit temporal feature engineering	Structured tabular data; interpretability requirements; limited compute resources
XGBoost	Ensemble (Boosting)	Strong regularisation; high accuracy on engineered features; sequential error correction; fast training	Temporal patterns require manual feature engineering; may overfit on minority fault classes without class weighting	Large feature sets; class-weighted imbalanced data; competition-grade baselines
LSTM	Recurrent Neural Network	Captures long-range temporal dependencies natively; learns from raw sequences without manual feature engineering; suitable for multivariate telemetry	Data-hungry; requires GPU; low interpretability; sensitive to hyperparameters; prone to overfitting on small datasets	Long multivariate sequences; GPU available; training data > 50,000 observations
CNN-LSTM Hybrid	Convolutional + Recurrent	Local feature extraction (CNN) combined with long-range dependency modelling (LSTM); state-of-the-art on complex temporal data; captures both short anomalies and long degradation trends	High computational cost; complex hyperparameter space; less interpretable; requires more data than LSTM alone	High-dimensional multivariate series; complex temporal patterns; research benchmarking
TCN (Temporal Conv. Network)	Dilated Causal Convolution	Parallelisable training (unlike LSTM); stable gradients; effective for long sequences; competitive performance with LSTM	Less established than LSTM in EV fault literature; interpretability similar to LSTM	Long sequences with parallelisation constraints; GPU-efficient training contexts
Classical (ARIMA, CUSUM)	Statistical / Rule-based	Lightweight; interpretable; no labelled training data required	Cannot model multivariate interactions; assumes linearity; poor on complex fault signatures; consistently outperformed by ML approaches	Univariate drift detection; resource-constrained baselines; simple threshold-based alerting

Note. Performance ranges reflect macro-averaged F1-scores across included studies using each architecture. GPU = Graphics Processing Unit; TCN = Temporal Convolutional Network.

An important result is that there was no single architecture that had a definite advantage over the others in all the fault types and application scenarios. The idea of an architecture-by-context interaction prompts the methodological stance adopted by Serradilla et al. (2022b) for comparing architectures in specific operational contexts.

4.3 Usage Stratification Evidence (RQ2)

The most important dimension in the EV predictive maintenance literature that is under explored is mentioned in this review and that is usage stratification. The 77 studies included – the main empirical contribution of this review – only four (5.2%) conducted any usage-stratified analysis. Theiseller et al. (2021) acknowledged that their experimental results were confounded by operational heterogeneity, but did not report the results separately for the various usage profiles and did not explicitly mention this as a study limitation. Serradilla et al. (2022b) state that two major structural challenges to the seamless application of predictive maintenance models are fleet heterogeneity, and operational subgroups where the performance of the predictive model based on the fleet of heterogeneous data cannot be consistently assessed.

This is an appreciable deficit in practice. A model that works well for a combined pooled set of heterogeneous vehicles can perform much worse on the highest risk, highest cost “heavy use” commercial vehicles, the vehicles for which the failures of the model have the greatest effect. Fleet operators who are using aggregate performance statistics don't know if a deployed model is reliable for their particular operational profile. Earlier, Ibrahim et al. (2025) gave the experimental and modelling study of the operation of electric buses and proved the usage-stratified analysis from the empirical point of view by showing that high duty usage can result in substantial degradation of the battery when compared to moderate duty usage.

None of the 77 studies used in this review systematically compared the performance of models from 4 or more different usage intensity categories simultaneously and no study investigated whether the most predictive variables in the telemetry used to forecast DTCs varied by

usage profile. This is the biggest justification for designing the accompanying empirical study.

4.4 Sensor-Level Precursor Patterns (RQ3)

A consistent set of telemetry variables were identified across all of the included studies as being systematic, in advance of DTC events. Temperature of the battery and temperature change rate were determined as significant predictors in 68.2% of the battery fault studies (n = 30/44 battery studies). Of the 24 battery studies, 54.5% used state of charge (SOC) instability (variance in SOC or rate of change within window). In motor related fault studies, motor temperature and RPM anomalies were the most prevalent. Among the studies that discussed charging system DTCs, charging voltage irregularity was the largest contributor to the DTCs.

Previous research by Zhao et al. (2019) showed that the cross-variable features produced by multiplying the magnitude of the current with battery temperature (as a cross-variable) were better than any single-variable threshold for prediction of the thermal fault. When compared with using only raw sensor values, rolling statistical features (mean, standard deviation and derivatives on battery rates of change computed over 6-hour and 24-hour windows) found by Zhang et al. (2019) allowed for a 12 percentage point increase in battery fault classification accuracy. The results support directly multivariate feature engineering as the data preparation method used in the subsequent empirical study.

Most importantly, none of the studies reviewed in this paper compared usage of sensor precursors under heterogeneous usage profiles using the same analytical framework. It is an empirical question whether the most predictive sensor variables for the DTC forecasting, vary according to the operational profile or not.

4.5 DTC Categorical Forecasting (RQ1 and RQ2)

The vast majority of included studies, 61 of 77 (79.2%) treated fault prediction as a binary classification problem (fault vs. no-fault) or as a regression task (remaining useful life estimation). Only 16 studies (20.8%) modelled fault prediction as a multi-class categorical DTC forecasting problem. Theissler et al. (2021) demonstrated that individual DTC categories exhibit systematically distinct temporal precursor signatures in sensor data, providing empirical justification for the categorical approach, and observed that different fault types showed distinct lead-time and variable importance profiles within their dataset.

Among the 16 multi-class studies, macro-averaged F1-scores ranged from 0.71 to 0.91, with higher performance generally associated with datasets exceeding 100,000 observations and with architectures that explicitly modelled temporal dependencies (LSTM, CNN-LSTM). All studies achieving macro-F1 above 0.83 reported the use of class imbalance handling techniques — class-weighted loss functions, focal loss objectives, or synthetic minority oversampling. Multi-class DTC categorical forecasting has significant operational benefit compared to binary prediction: A specific DTC category alert within 48 hours helps with preparedness of specific components and parts which is helpful for maintenance planning, a generic DTC alert does not offer any specificity.

4.6 Emerging Market Applicability (RQ4)

There is limited to no convincing evidence of how the use of such predictive maintenance frameworks can be applied to data-poor new EV markets. Of the 77 studies in total, only two specifically mentioned emerging market contexts of EV deployment, with none offering empirical evidence of the model transfer from a mature market model to the emerging market operational context. Data scarcity is also a major structural factor that hinders the adoption of ML-based predictive maintenance in new fleet contexts, as deep learning approaches require large amounts of high-quality, labelled data, which are not readily available in new and emerging fleet contexts, as reported by Serradilla et al. (2022b). They review some of the theoretical solutions that seem appropriate to their problem, which include transfer learning and domain adaptation, but encounter no empirical evidence in the context of EV maintenance.

No studies have been included that address the infrastructure and operational context constraints unique to Kenya. The Draft National E-Mobility Policy 2024 (Ministry of Roads and Transport, Kenya) shows that there is a current lack of technical capacity in Kenya for EV diagnostics and data science. In contrast to the highway driving cycles typical of most European and North American training datasets, BasiGo electric buses (BasiGo, 2025) running on high-frequency urban lines in Nairobi operate with stop-start driving cycles. The warm environmental temperature (25–30°C) in Kenya poses unique battery thermal management issues, which are not captured in temperate climate datasets. These context differences highlight the need for explicit validation of externally-trained models prior to operational deployment in this context.

5. Discussion

5.1 Principal Findings

There are five main findings from a systematic review of 77 peer-reviewed studies. The performances reported in sequential deep learning architectures (particularly CNN-LSTM hybrids) were generally higher than the performances reported for ensemble tree-based methods and classical statistical baselines for the EV fault forecasting tasks, with a median macro-averaged F1 score of 0.86 for CNN-LSTM hybrids; however, as noted in Section 4.2, no single architecture induced superior performance in all fault categories and usage contexts, and reported performances varied significantly with data characteristics, preprocessing, hyperparameter tuning, class imbalance handling, and prediction horizon, so this result should be taken as a general trend rather than a standard one.

Second, and most importantly, there was almost no stratification of usage analysis (which was included in 5.2% of studies), which is the most serious methodological limitation of the existing literature.

Third, the use of multi-class DTC categorical forecasting is still not widely used compared to its practical usefulness (79.2% of the studies used binary categorical forecasting).

Fourth, there is a lack of characterisation of the distribution of lead-time which is nearly absent, thereby leaving out the possibility for an operational feasibility assessment. Fifth, no study delivers empirically sound information on how to overcome the barrier of data scarcity in markets where EV fleets are new and emerging, like Kenya.

Fifth, Serradilla et al. (2022b) conclude that the lack of data is the main structural hurdle to predictive maintenance implementation in new and emerging EV fleet contexts, but there is no study with empirical information for overcoming this obstacle in markets like Kenya.

5.2 Comparison with Prior Reviews

This is the first systematic review to examine the intersection of ML-based DTC forecasting, heterogeneous usage stratification, and emerging market applicability in EVs. Theissler et al. (2021) provided the most comprehensive prior automotive predictive maintenance review but were restricted to the binary fault classification paradigm and did not examine usage stratification or emerging market considerations. Lee et al. (2014) and Jardine et al. (2006) addressed predictive maintenance broadly without EV specificity. Serradilla et al. (2022b) conducted the most thorough review of deep learning architectures for PdM, covering 156 papers,

but did not address EV-specific DTC forecasting, usage stratification, or emerging market deployment. The present review extends and updates the evidence base in three substantive dimensions: it focuses on EV-specific DTC forecasting, identifies usage stratification as a critical missing analytical dimension, and assesses evidence for data-scarce emerging market deployment contexts.

5.3 Research Gaps and Study Positioning

Based on the synthesis of evidence, four substantive research gaps are identified. These gaps are presented in Table 8 alongside their positioning in relation to the accompanying empirical study

Table 8. *Research Gaps Identified, Evidence Basis, and Contributions of the Accompanying Study*

Gap Identified	Limitation in Prior Research	How This Review Addresses It	Expected Contribution
Gap 1: Absence of usage-stratified performance analysis	Studies aggregate model performance across heterogeneous fleets, masking the effect of usage intensity on DTC occurrence rates and model accuracy	Quantifies that only 5.2% of studies stratified by operational profile; synthesises the limited evidence on usage-specific performance; calls for usage-stratified evaluation as a minimum reporting standard	Enables fleet operators to assess model applicability to their specific profile; provides empirical basis for rejecting aggregate-only performance metrics in deployment decisions
Gap 2: Dominance of binary fault classification over multi-class DTC forecasting	79.2% of included studies treat fault prediction as binary (fault/no-fault) or as regression (RUL estimation), despite multi-class categorical DTC forecasting being operationally more actionable	Synthesises the 20.8% of multi-class studies; documents performance range and architectural choices associated with categorical DTC forecasting; establishes multi-class as the methodologically superior paradigm for fleet operations	Motivates targeted maintenance preparation; enables fault-category-specific lead-time analysis; directly motivates the accompanying empirical study's multi-class formulation
Gap 3: Absence of lead-time distribution characterisation	Studies report point estimates (e.g., '48-hour horizon') without characterising the distribution of detection lead times, preventing operational feasibility assessment	Synthesises available lead-time evidence; identifies near-total absence of percentile-level reporting; articulates why distributional characterisation is operationally necessary for parts procurement and scheduling	Supports maintenance scheduling decisions; quantifies the proportion of faults detectable with sufficient advance warning to be operationally actionable
Gap 4: Absence of methodological guidance for data-scarce emerging market contexts	Predictive maintenance literature assumes data-mature operational histories; Serradilla et al. (2022) identify data scarcity as a primary barrier; no empirically validated guidance exists for Kenya or East Africa	Evaluates evidence on transfer learning and external-dataset approaches; identifies data scarcity as a structural barrier; articulates transferability requirements for methodology deployment in new fleet contexts	Provides a pragmatic pathway for EV fleet operators in Kenya and similar markets to access predictive maintenance capability without waiting for local data maturity

Note. All four gaps are grounded in quantitative evidence from the 77 included studies and in findings from verified peer-reviewed sources. Each gap directly motivates a specific design decision in the accompanying empirical study.

Gap 1 — Usage-Stratified Analysis: The most significant gap identified is the lack of usage-stratified performance analysis. If a fleet has some vehicles performing at a high intensity level and others at a low one, then the operators can't responsibly use a predictive model based on the average of their performance statistics. Ibrahim et al. (2025) showed directly from empirical data that there is a systematic difference in the occurrence rates of DTCs and the degradation trends between high-intensity and moderate-duty electric bus operations, lending empirical support to the viewpoint that there are systematic differences in the incidence rates of DTCs and the degradation trends between the operational profiles. Serradilla et al. (2022b) also show that fleet heterogeneity is one of the most significant structural barriers to model deployment and average performance metrics can lead to wrong deployment decisions. The use of experimental designs stratified by usage should be used as a standard requirement in future studies.

Gap 2 – Multi-Class DTC Forecasting: Binary fault classification oversimplifies the operational benefits that predictive maintenance systems can provide. Multi-class categorical DTC forecasting proved to be more operationally useful in 20.8% of studies, evidenced by the ability to predict the fault category-specific lead time. An expansion of multi-class DTC forecasting in conjunction with usage stratification is the methodological pathway that is most likely to yield operationally deployable frameworks.

Gap 3 — Lead-Time Distribution Characterisation: Point estimates of prediction horizons are not enough for operational planning. Fleet operators need to know what percentage of faults they can identify and buy parts for, or schedule their staff for, in advance in order to plan vehicle rotations. These distribution reporting should be considered as a baseline for future studies in EV fault forecasting.

The current lack of empirically-backed guidance to implement well-developed predictive maintenance practices from the mature markets to the environment of limited data in an emerging market like Kenya is a practical constraint to accessing EV maintenance technology fairly. First commercial EV deployments in East Africa started in 2022. Serradilla et al. (2022b) present transfer learning as the most promising approach to mitigate the scarcity of data in such contexts, which has yet to be validated empirically in an EV maintenance context. There is a need for systematic methodological templates, based on well-developed external data sets.

5.4 Implications for Practice

Aggregate predictive maintenance performance measures should be taken with a pinch of salt by fleet operators. An overall good F1 score

on a pooled dataset could dramatically fail to perform well on the actual usage profile that forms the basis of their fleet. Until performance evidence for usage come, operators need to try to find a model validated in data that is representative of their operational environment.

This review suggests that there is not yet an empirically validated predictive maintenance framework that currently captures the context in which fleet operators in Kenya operate, such as BasiGo electric buses, Roam electric motorcycles or the new last mile delivery companies. Given the state of the existing external data source, the usage- stratified modelling approach outlined in the accompanying empirical study offers a pragmatic and evidence-based route forward which is intended to set up a local validation phase.

5.5 Implications for Future Research

This review identifies 5 research priorities. Future studies should: (i) report using usage-stratified experimental designs and report precision, recall and F1-score separately for distinct usage profiles, with statistical significance testing of between-profile differences; (ii) shift from binary fault classification to multi-class DTC categorical forecasting, as the primary prediction target; (iii) report lead-time distributions at the percentile level for all the fault categories and usage profiles; (iv) empirically validate the transfer learning approaches proposed in this review by Serradilla et al. (2022b) for emerging market EV fleet contexts; and (v) develop standardised reporting guidelines for EV predictive maintenance studies — similar to the CONSORT guidelines in clinical research — that would allow for systematic comparisons across studies and reduce the large methodological heterogeneity that hindered a meta-analysis in this review.

5.6 Strengths and Limitations

The five strengths of this review are: a comprehensive search, using four databases, performed in March–April 2026, which included all literature through April 2026; two independent reviewers performed the search and confirmed a high level of inter-rater agreement (kappa: 0.81 and 0.87); the reviewers used structured, quality assessment criteria covering five bias domains for all 77 reviewed studies; and a thematic narrative synthesis was performed and the evidence was clearly mapped to research questions and gaps identified. Limitations include the restriction to English language publications, which may have resulted in the exclusion of relevant work published in other languages; the near absence of usage stratification in literature, which precluded quantitative meta-analysis; the inherent limitation of systematic reviews, that is, they focus on the literature that has been published to date, reflecting a gap in research priorities that usage intensity is important, rather than a lack of research interest.

The other weakness is the use of an evidentiary basis for the review's conclusions in the case of Kenya and other emerging African EV markets. The data from only 2 of the 77 studies (2.6%) came from operational settings in Africa, and these two were not from East Africa. The review also repeatedly mentions Kenya as a motivating case for the empirical study in Section 3, but the conclusions made about the applicability of the emerging markets are largely inferential — based on a thin direct evidence base and are drawn from studies conducted in data-rich markets, as well as context on how each of these models are implemented in Kenyan (or East African) fleet operations (e.g., BasiGo deployments, Roam deployments) and the Draft National E-Mobility Policy. The claims made in the review are specific to Kenya and are intended to highlight the research gap and a potential reason for the empirical study that follows, rather than to provide confirmation of the performance of the models in an African EV fleet.

Three additional sources of bias should be noted. First, there is likely publication bias because studies reporting weak or null predictive performance are less likely to be submitted or accepted for publication in this review, thus increasing the apparent effectiveness of the ML architectures synthesized. Second, database bias can occur because the search was limited to IEEE Xplore, ScienceDirect, ACM Digital Library and Google Scholar; other regional or discipline-specific databases (e.g. automotive engineering databases other than these) and grey literature (e.g. industry technical reports, conference posters, theses) were not searched and were explicitly excluded by the eligibility criteria. This could have led to an undue bias towards applied or industry-based evidence, such as work from emerging markets, which is less likely to be found in the indexed academic literature. Third, language bias, due to the English language restriction mentioned above, adds to the database bias because it may have filtered out Chinese-language research at a time when the majority of the studies included in the database reported Chinese EV deployments. These three biases taken together indicate that the evidence base of the review should be interpreted as a representative sample of the published, English-language, database-indexed literature and not as a complete picture of the extent of evidence available in this field.

Subgroup meta-analysis/meta-regression was considered but deemed not feasible due to the overall metric heterogeneity. An informal, exploratory subgroup meta-analysis was carried out on studies that reported the macro-averaged F1-score for LSTM-based architectures, but even within this subgroup the studies differed greatly in sample size, class imbalances, prediction horizon, and train-test splitting method (see Table 5), all of which could introduce confounding in a pooled effect estimate. Studies reporting a common metric (F1-score) with enough study-level moderator data (with fewer than 20 studies for most candidate moderators) were also considered, but the number of studies reporting a common metric was determined to be too small in relation to the number of candidate moderators to produce a stable or interpretable model when using meta-regression. Due to these factors, the narrative synthesis approach was kept as the more defensible approach to the analysis, in line with the SWiM guidance (Campbell et al., 2020); in the future, with a larger and more metric consistent evidence base, subgroup meta-analysis may be attempted.

In addition to the formal quantitative synthesis, a pattern observed qualitatively across the studies included in the synthesis is also briefly commented upon: the larger the datasets included in the studies, the higher the macro-averaged F1-scores reported, and the higher the macro-averaged F1-scores, the higher the number of studies that explicitly model temporal dependency (models featuring LSTM, CNN-LSTM). Smaller datasets (up to about 20,000 observations) more often yielded F1-scores under 0.75, notably without class-imbalancing methods. This trend is in line with, but does not necessarily imply, the overall intuition that deep sequential architectures need to be trained with larger amounts of data to outperform ensemble or classical approaches in terms of representation. However, the apparent association of dataset size with other study characteristics (e.g., publication year, class-imbalance handling, train-test splitting on time) found in the included literature must be interpreted with caution, as it may not represent a true causal or independent effect of dataset size only, but rather a reflection of other study characteristics that were not accounted for.

6. Conclusion

The systematic review was conducted based on 77 peer-reviewed papers published between January 2015 and April 2026, following the PRISMA 2020 guidelines, addressing the field of diagnostic fault forecasting in EVs using machine learning techniques. The review shows that although the technical feasibility and economic motivation of this approach are firmly established, the CNN-LSTM hybrid architectures presented yield up to a macro-averaged F1-scores of

0.92 under the controlled experimental conditions. The literature is methodologically immature, however, on four key aspects that significantly hinder the use in practical, real-life heterogeneous fleet operations.

Because very little usage stratified performance analysis was provided in the included studies (5.2%), fleet operators cannot determine if performance models published in the literature are applicable to their operational profile. Most studies result in a binary classification of fault (79.2%) while only a few result in multi-class DTC categorical forecasting, which makes most of the models lack the necessary specificity to take any specific action in the preparation for maintenance. Without the lead-time distribution characterisation, the assessment of operational feasibility is not possible. But there is no empirically proven methodology guidance for data-scarce emerging EV markets – including Kenya, which started commercial EV deployments in 2022 – so operators in these markets do not have any evidence-based methodology to guide deployment decisions.

The four gaps stimulate and justify the related empirical study that will design and benchmark machine learning models for four different usage intensity profiles (UIPs) for diagnostic fault forecasting on a five-year, 172,800-observation telemetry dataset of EVs. The empirical study will directly contribute to the usage stratification literature, the multi-class DTC forecasting literature, lead-time distribution characterisation, and methodological guidance for emerging markets, such as Kenya.

References

Africa E-Mobility Alliance. (2025). Africa e-mobility report 2025. AfEMA. Retrieved April 2026, from https://africaema.org/africas-electric mobility-status-2025/

BasiGo. (2025, August 4). BasiGo deploys 100th electric bus in East Africa. CleanTechnica. Retrieved April 2026, from https://cleantechnica.com/2025/08/04/basigo-deploys-100th-electric-bus-in-east-africa/

Campbell, M., McKenzie, J. E., Sowden, A., Katikireddi, S. V., Brennan, S. E., Ellis, S., Hartmann-Boyce, J., Ryan, R., Shepperd, S., Thomas, J., Welch, V., & Reeves, B. C. (2020). Synthesis without meta-analysis (SWiM) in systematic reviews: Reporting guideline. BMJ, 368, l6890. https://doi.org/10.1136/bmj.l6890

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hossain, M., Nasser, M., & Ahmad, S. (2023). Internet of Things-based smart vehicle health monitoring and predictive maintenance system. IEEE Access, 11, 71008–71023. https://doi.org/10.1109/ACCESS.2023.3292040

Hu, X., Xu, L., Lin, X., & Pecht, M. (2020). Battery lifetime prognostics. Joule, 4(2), 310–346. https://doi.org/10.1016/j.joule.2019.11.018

Ibrahim, H., Ali, A. M., & Attia, T. (2025). Optimizing electric bus performance via predictive maintenance: A combined experimental and modeling approach. Frontiers in Future Transportation, 5, Article 1506866. https://doi.org/10.3389/ffutr.2024.1506866

International Energy Agency. (2024). Global EV outlook 2024: Trends and developments in electric vehicle markets. IEA Publications. https://www.iea.org/reports/global-ev-outlook-2024

Ismail, S. Y., Hussain, Z. S., & Shanshal, A. K. (2025). Intelligent predictive maintenance for urban energy and transportation systems: A hybrid AI approach. NTU Journal of Renewable Energy, 8(1), 39–50. https://doi.org/10.28991/NTU-JRE-2025-08-039

Jardine, A. K., Lin, D., & Banjevic, D. (2006). A review on machinery diagnostics and prognostics implementing condition-based maintenance. Mechanical Systems and Signal Processing, 20(7), 1483–1510. https://doi.org/10.1016/j.ymssp.2005.09.012

Lee, J., Wu, F., Zhao, W., Ghaffari, M., Liao, L., & Siegel, D. (2014). Prognostics and health management design for rotary machinery systems: Reviews, methodology and applications. Mechanical Systems and Signal Processing, 42(1–2), 314–334. https://doi.org/10.1016/j.ymssp.2013.06.021

Ministry of Roads and Transport, Kenya. (2024). Draft national e-mobility policy framework 2024–2030. Government of Kenya, Nairobi.

Moubray, J. (1997). Reliability-centered maintenance (2nd ed.). Industrial Press.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, n71. https://doi.org/10.1136/bmj.n71

Qin, Y., Wang, X., & Zou, J. (2021). Fault diagnosis of electric vehicle motor system based on deep learning. IEEE Transactions on Vehicular Technology, 70(9), 8619–8629. https://doi.org/10.1109/TVT.2021.3096888

SAE International. (2018). Surface vehicle recommended practice: E/E diagnostic test modes (SAE Standard No. J1979-2). SAEInternational. https://doi.org/10.4271/J1979-2_201803

Serradilla, O., Zugasti, E., Ramirez de Okariz, J., Rodriguez, J., & Zurutuza, U. (2022a). Adaptable and explainable predictive maintenance: Semi-supervised and active learning approaches. Applied Sciences, 12(14), Article 7273. https://doi.org/10.3390/app12147273

Serradilla, O., Zugasti, E., Rodriguez, J., & Zurutuza, U. (2022b). Deep learning models for predictive maintenance: A survey, comparison, challenges and prospects. Applied Intelligence,52(10),10934–10964. https://doi.org/10.1007/s10489-021-03004-y

Theissler, A., Pérez-Velázquez, J., Kettelgerdes, M., & Elger, G. (2021). Predictive maintenance enabled by machine learning: Use cases and challenges in the automotive industry. Reliability Engineering & System Safety, 215, Article107864. https://doi.org/10.1016/j.ress.2021.107864

Zhang, W., Yang, D., & Wang, H. (2019). Data-driven methods for predictive maintenance of industrial equipment: A survey. IEEE Systems Journal, 13(3), 2213–2227. https://doi.org/10.1109/JSYST.2019.2905565

Zhao, R., Yan, R., Chen, Z., Mao, K., Wang, P., & Gao, R. X. (2019). Deep learning and its applications to machine health monitoring: A survey. Mechanical Systems and Signal Processing,115,213–237. https://doi.org/10.1016/j.ymssp.2018.05.05