vehicles

MDPI

Article

Multi-Domain Machine Learning Framework for Electric Vehicle Charging Prediction

Hanan Thwany [1], Muhammad Alolaiwy [1,2,*] and Mohamed Zohdy [1]

[1] School of Engineering and Computer Science, Oakland University, Rochester Hills, MI 48309, USA; hhthwany@oakland.edu (H.T.); zohdyma@oakland.edu (M.Z.)
[2] Department of Computer Science and Engineering, Yanbu Industrial College, Royal Commission for Jubail and Yanbu, Yanbu Industrial City 46451, Saudi Arabia
* Correspondence: alolaiwy@oakland.edu

Abstract

Electric vehicle (EV) adoption is rising rapidly, creating growing challenges for charging infrastructure planning, energy demand management, and grid stability. However, most existing studies rely on single-domain data, such as behavioral charging sessions or station metadata, which limits their ability to capture the joint effects of user behavior, charger characteristics, and market context. To address this gap, this study proposes a multi-domain machine learning framework for EV charger-type prediction by integrating behavioral, infrastructure, and market-level data. Behavioral charging logs are transformed into structured event-token sequences and modeled using XLM-RoBERTa (Cross-lingual Language Model–RoBERTa), which is used here as a transformer-based sequence encoder to capture long-range dependencies in charging behavior. Structured infrastructure and market features are modeled using LightGBM and TabNet. The study contributes a unified multi-domain framework, a systematic comparison of transformer and tabular-learning models, and a broader evaluation through ablation analysis, cross-validation, confusion matrix analysis, and confidence calibration. The results show that multi-domain fusion consistently improves performance over single-domain learning. XLM-RoBERTa achieved the best overall performance on the fused dataset, with 98.76% accuracy and 97.86% weighted F1-score, while TabNet demonstrated stronger calibration and deployment reliability.

Keywords: electric vehicle charging; charger type prediction; transformer models; XLM-RoBERTa multi-modal learning; data quality; smart charging infrastructure

Check for updates

Academic Editor: Weixiang Shen

Received: 29 January 2026
Revised: 7 May 2026
Accepted: 11 May 2026
Published: 20 May 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

The primary factors driving the global transformation to electric vehicles (EVs) are increased environmental awareness, regulations, and technological advancements [1]. Also, this shift is accompanied by several significant issues, notably those related to the administration and expansion of the charging infrastructure [2]. The trend in the transportation sector is the rapid rise in the sales of EVs [3]. They provide a clean, low-carbon, and environmentally sustainable alternative. Looking at these persistent trends, EVs are scheduled to become the new leaders in the market, replacing traditional internal combustion engine (ICE) vehicles [4]. According to estimates based on the Sustainable Development Scenario (SDS) from 2020 to 2030, the share of electric cars is expected to increase significantly over the coming year, reaching 13.4% by 2030 [5]. Expanding the use of electric vehicles also provides several vital advantages: it improves energy security by utilizing diversified

energy sources; it stimulates the economy by creating new, innovative industries; and, most importantly, it is environmentally sustainable and helps in the fight against climate change because it emits less pollution and is more renewable [6]. Although electric cars have many benefits over internal combustion engine cars, the power grid can still face problems if EVs are used on a large scale [7]. The unsteady and changing character of EVs' charging requirements may result in greater peak loads, more fluctuations in voltage and frequency, and an increase in total energy consumption, which could ultimately limit the grid's stability [8,9].

Thus, predictive analytics becomes essential as it allows for the prediction of charging needs, the avoidance of overloaded power grids, and the guarantee of reliable and efficient EV charging services [10,11]. The expansion of electric vehicle markets necessitates quicker development and better service of the charging infrastructure, which is the focus of this study. Accurately predicting the consumers' charging nature will assist in mitigating the issues arising from the peak energy demand, thus lowering the energy management costs and making the users more satisfied [12]. Therefore, charging electric vehicles with the help of slot analytics is the key to the future of sustainable transportation.

On the one hand, the adoption of innovative smart-grid technologies has played an essential role in not only modern but also future energy systems. On the other hand, with the rising number of IoT devices, the forecasting accuracy of electric vehicle (EV) charging demand has become a requirement rather than an option [10,13]. Smart meters and real-time monitoring, along with two-way communication between EVs and charging infrastructure, can make it possible to use more flexible and adaptable load management strategies. These strategies will thus not only reduce the risk of grid overload but also increase the efficiency of energy distribution [14,15]. Furthermore, the incorporation of clean resources, such as solar and wind, into the charging network has led to a series of difficulties in supply–demand balancing due to the variability in supply and unpredictability in user demand [16,17]. To address these challenges, machine learning frameworks should be highly precise, capable of handling different data formats without problems, adaptable in real-time, and scalable on a regional basis in different localities.

There are several major challenges that electric vehicle charging-type prediction through modeling still need to be overcome, despite its high potential [18]. Some of these are the variety of datasets involved, a lack of consistency in data collection methods, and human behavior, which is hard to predict [19]. Problems like missing data or incomplete data, no standard metadata for charging stations, and the models requiring a powerful computer to run are some of the limitations that further complicate the situation and make the task of accurate prediction modeling a challenging one.

The work is mainly about greatly extending the current literature through an integrative methodological approach that not only combines but also makes the most of the varied global datasets gotten from different sources. The use of transformer-based models, especially the very well-tuned XLM-RoBERTa, has resulted in us achieving a much better level of predictive accuracy and reliability than the traditional models. This study evaluation is quite comprehensive in that it quantitatively expresses the most important properties of accuracy, precision, recall, and F1-score. It provides a detailed confusion matrix analysis, which helps pinpoint exactly where the model is making errors for each class. Such methodological rigor opens up avenues for more informed strategic decisions by stakeholders responsible for EV infrastructure planning and operations.

In this article, merging different types of EV data is considered not merely an end in itself but a way to resolve one precisely identified forecasting issue: EV charger-type prediction. The reason for mixing behavioral, infrastructure, and market-related info is to check if a single combined representation from various domains can not only help to

improve the classification performance but also make the model stronger and more capable of working in different operating contexts. Creating the research around one forecast target only allows the suggested setup to go beyond just tweaking the data and give an evaluation from the perspective of the use of heterogeneous EV data sources in helping to classify charger types.

The following are the contributions of the study:

- This study evaluates three fundamentally different model families, XLM-RoBERTa, LightGBM, and TabNet, across three EV-related data contexts, namely Behavioral Patterns, Infrastructure Data, and Market Dynamics.
- This study adopts a two-stage experimental design in which each model is first trained and evaluated separately on each individual dataset, and then re-evaluated on the merged multi-source dataset to examine the effect of cross-domain data integration on predictive performance.
- This study analyzes cross-domain robustness by comparing model behavior across heterogeneous data modalities and structures, thereby highlighting the trade-offs among accuracy, generalization, and deployment-oriented reliability.

The remaining parts of the article are structured below: Section 2 describes the literature review, Section 3 describes the proposed methodology, Section 4 contains the results and discussion, and Section 5 contains the conclusion and future work.

2. Literature Review

Studies on electric vehicle (EV) charging behavior modeling using various machine learning methods, including traditional and advanced ones, have significantly increased over the last few years.

Early methods were primarily based on conventional time-series models, such as Autoregressive Integrated Moving Average (ARIMA), Seasonal Autoregressive Integrated Moving Average (SARIMA), and Seasonal Autoregressive Integrated Moving Average with Exogenous Variables (SARIMAX), which were often compared with machine learning approaches [20]. Illustrated with the use of ARIMA, baseline load forecasting is feasible. In contrast, Hitesh et al. [16] presented SARIMAX plus Convolutional Neural Networks (CNNs), Gated Recurrent Units (GRUs), and Extreme Gradient Boosting (XGBoost), thus proving that hybrid models can be better than statistical models alone in monthly EV charging demand prediction. Studies have increasingly adopted tree-based ensemble methods and explainability tools such as SHAP. One significant study by T. Zhang et al. [17] applied XGBoost, Light Gradient Boosting Machine (LightGBM), CatBoost, Random Forest, and Linear Regression to a California EV dataset. They found that XGBoost was the most accurate model and then used Shapley Additive explanations (SHAP) to determine the influence of variables such as renewable penetration and grid stability. The complexity of spatiotemporal charging data demanded adjustments to the neural methods. Hüttel et al. [18] demonstrated in 2021 a Temporal Graph Convolutional Network (T-GCN) that represented spatial relations between stations not only for short-term but also for long-term demand prediction.

In another study, Aduama et al. [19] demonstrated a weather-aware Long Short-Term Memory (LSTM) model that achieved 3.29% higher accuracy through multi-feature data fusion for station load forecasting. Recently, transformer architectures have also been experimented with. In [21], research by F. Zeng et al. was published, which suggested a federated, privacy-preserving training and truncated attention transformer model for making intelligent charge duration predictions. In [22], Manzoor et al. conducted a study comparing transformer attention mechanisms for EV charging demand forecasting and confirmed that transformers outperform traditional and recurrent architectures. Li et al. [23] presented a

probabilistic forecast model of power for EV charging stations, which was a fusion of LSTM and Proximal Policy Optimization (PPO)-based reinforcement learning for addressing the uncertainties in charging demand, user behavior, and renewable-energy availability.

The evolution from conventional time-series and machine learning models to transformer-based architectures indicates the increasing demand to capture intricate temporal interactions and spatial variability in electric vehicle charging data [4]. However, current transformer-based studies remain relatively narrow in focus. They have predominantly dealt with single-location datasets or only one type of prediction, and have also mostly not combined different global datasets.

Previous research on electric vehicle (EV) charging analytics has mainly looked at either behavioral or infrastructure data, using traditional machine learning or deep learning methods. However, most of these studies remain based on a single dataset, lack adequate data quality control, and provide minimal integration of different types of EV-related information. Problems like label inconsistency, data leakage, and very few cross-domain evaluations make the practicality and reliability of many published results questionable. In order to overcome these drawbacks, this paper introduces a unified multi-domain approach to EV charger-type prediction, which not only integrates behavioral, infrastructure, and market data but also operates within a standardized preprocessing and evaluation framework. Along with strict data preparation, the paper also carries out a comparative evaluation of different model families for achieving a better and more practically meaningful understanding of the predictive performance in various heterogeneous EV data contexts.

3. Materials and Methods

To improve methodological clarity and reproducibility, the Materials and Methods Section is organized into a structured workflow consisting of data collection, data preprocessing, feature engineering, spatial dataset integration, model development, training strategy, validation procedures, and performance evaluation. This sequential structure ensures that each stage of the proposed framework is clearly described and logically connected.

The proposed method of electric vehicle (EV) charger-type prediction via multi-modal machine learning consists of a few essential steps. Figure 1 shows the overall workflow of the multi-domain machine learning framework that helps in predicting the type of an electric vehicle (EV) charger. This framework brings together different kinds of data sets that represent three different and complementary parts of the EV charging ecosystem: user charging behavior, charging infrastructure, and market environment.

User behavior data records the details of an EV user's interaction with a charging system, which over time changes in charging station usage, session duration, and amount of energy consumed. Info about the charging infrastructure is about the physical attributes of charging terminals, such as their location, type of charger, capacity of the connector, and operational status. Market dynamics data supply the necessary information on how the availability, spread, and growth of EV charging mechanisms are happening at a regional level. When combined, these datasets give a well-rounded picture of the EV charging scene.

Once the data are collected, the sets are taken through preprocessing and feature engineering to improve their consistency and reliability. Some of the things done at this stage are the treatment of missing values, elimination of duplicate records, conversion of categorical features to numbers, and scaling of numeric variables. Features related to time, like hour of the day, day of the week, and duration of the charging, are obtained from the timestamps of the charging sessions. Apart from that, features like charging speed and electricity consumption habits are created to add to the power of prediction. Techniques for

identifying outliers, like removing data points outside the quartiles, are used to eliminate the value that deviates so much that it can harm the model training.

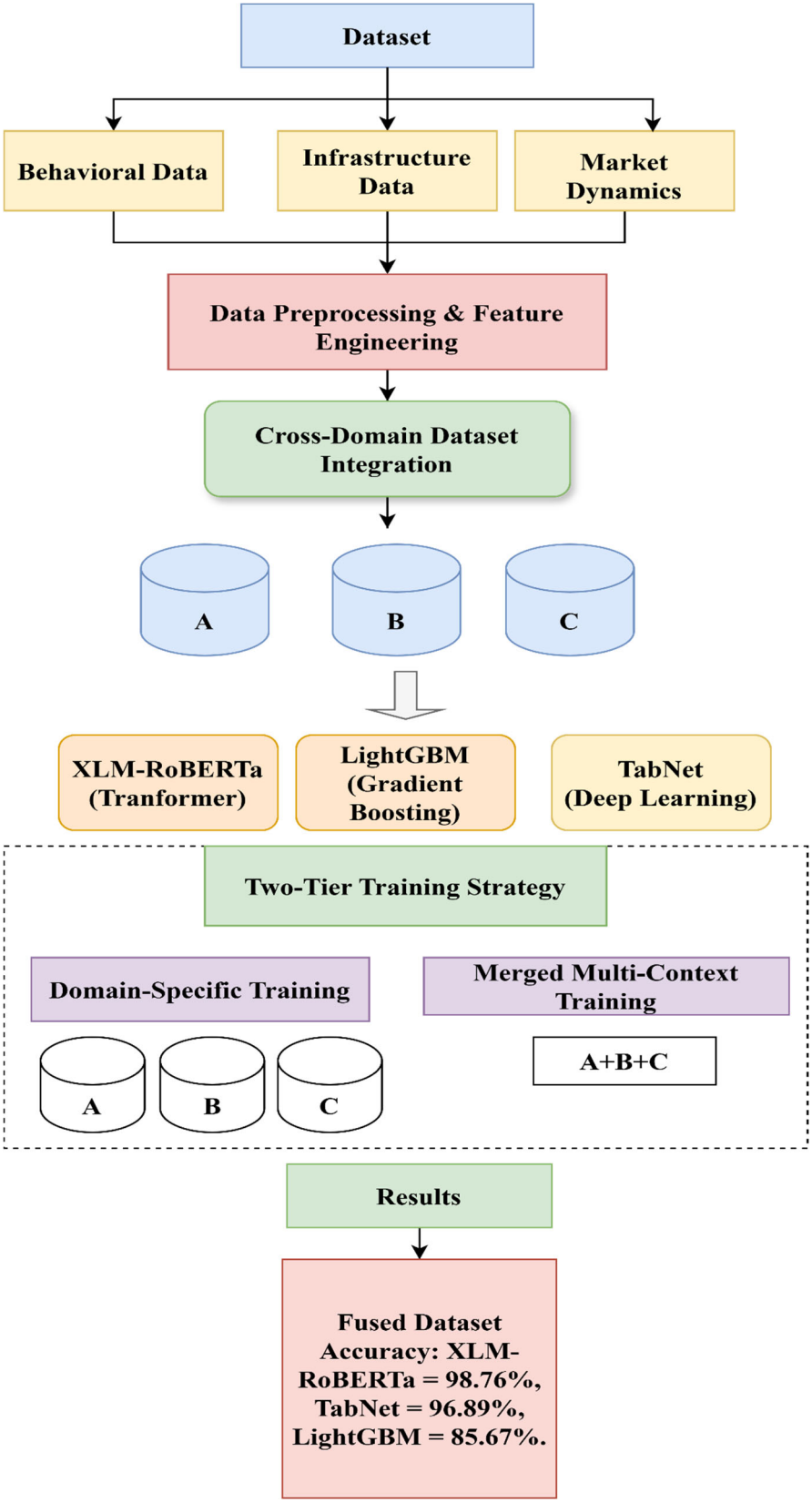


Figure 1. The proposed architecture diagram.

The next step after preprocessing is dataset integration of behavioral, infrastructure, and market data through a cross-domain approach. This process leads to the alignment of heterogeneous datasets and, subsequently, a unified multi-domain dataset is created, and the domains are referred to as A (Behavioral Data), B (Infrastructure Data), and C (Market Dynamics). By combining the datasets, the learning algorithms can explore the connections between user behavior, charging infrastructure properties, and market environment. After that, it is the turn of the processed data to be fed into three distinct machine learning models: XLM-RoBERTa (Cross-lingual Language Model, RoBERTa), LightGBM (Light Gradient Boosting Machine), and TabNet (Attentive Interpretable Tabular Neural Network). Among these, XLM-RoBERTa is employed for modeling behavioral charging sequences using transformer-based attention mechanisms that are capable of capturing long-range temporal dependencies. On the other hand, LightGBM is utilized for structured infrastructure features through gradient boosting decision trees that are specifically optimized for tabular data. Meanwhile, TabNet is based on a deep learning model with sequential attention mechanisms that allow not only feature selection but also deep interaction modeling in tabular datasets.

First of all, a series of training experiments is run to understand the depth of learning the models can achieve. At first, domain-specific training, where each dataset (A, B, and C) is used individually to train the models, is done. This is aimed at finding out the effectiveness of the model within the isolated domains. Then the two-stage training is carried out, in which the merged multi-context train consists of the combined dataset (A + B + C). Thanks to this, the models can learn from different domains and thus be better at generalization. Lastly, there are multiple performance indicators by which the trained models are assessed. Apart from classification metrics, the analyses of the confusion matrix and calibration are used for cross-checking. Together, the evaluation processes reveal different aspects of the model, such as its accuracy, prediction behavior at the class level, and confidence reliability. In the end, the models with these capabilities offer the prediction of EV charger types with a level of robustness that supports intelligent charging infrastructure planning and management.

3.1. Dataset Description

The charging station used for this research is fully equipped with the following data: EV charging session information, charging station metadata, and user behavioral patterns, see Table 1.

Table 1. Dataset description.

Name and Reference	Type	Number of Records	Features	Key Features
EV Charging Forecasting Base [24]	Charging sessions	2274	34	Session IDs, timestamps, energy, duration, user/vehicle, pricing inputs
Electric Vehicle Charging Patterns [25]	Charging sessions	~1320 sessions	~20	Battery capacity, energy consumed, duration, distance, cat'd: vehicle, station, user type.
Global EV Charging Stations Dataset [26]	Station-level	~5000 stations	~8–15	Location (lat/long), station ID, charger types/capacity, status, address/operator

3.1.1. EV Charging Forecasting Base

This dataset covers electric vehicle (EV) charging sessions and was primarily aimed at building and validating forecasting models. It is a record of highly minute-by-minute session parameters, e.g., the exact session connection and disconnection times, the overall duration of the charge, and the energy consumed per session (kWh). Besides serving as a great source for investigating EV charge consumption behaviors, the data also support the creation of prediction models and the implementation of smart charging infrastructure schedules and operations. This section introduces the EV Charging Forecasting Base dataset [24], hereafter referred to as Dataset 1 in Table 1.

3.1.2. Electric Vehicle Charging Patterns

This part of the paper explains in detail the Electric Vehicle Charging Patterns dataset [25], which is Dataset 2 in Table 1. The Electric Vehicle Charging Patterns dataset studies user and usage patterns of EV charging sessions. The dataset comprises around 1320 charging sessions, with very detailed records of session start and end times, charging duration, and total energy consumption. Furthermore, by temporal factors such as the time of day, peak usage periods, and the location of charging, this dataset isolates charging habits. It assists in examining charging habits, monitoring utilization trends, and even making administrative decisions concerning energy distribution and infrastructure location.

3.1.3. Global EV Charging Stations Dataset

This section zooms in on the Global EV Charging Stations dataset [26], which is Dataset 3 in Table 1. The Global EV Charging Stations dataset is a compilation of information regarding the locations of around 5000 electric vehicle charging stations all over the world. Each station's data covers the location (latitude and longitude), address, city, and country. The dataset mostly gives the technical details of the stations, capacity, connector type, and charging levels, along with an operational part that shows the status (active or inactive) of the station. The dataset can be used for a geo-analysis, planning the sustainable development of the infrastructure, and research questions about the availability and distribution of EV charging facilities worldwide.

This table summarizes the three datasets used in this study: EV Charging Forecasting Base (Behavioral Patterns, A) [24], Electric Vehicle Charging Patterns (Behavioral Patterns, A) [25], and Global EV Charging Stations (Infrastructure/Market, B–C) [26].

To address the issue of class imbalance, oversampling (SMOTE) is done only on the training split. The validation/test distributions are kept intact to reflect the real-world prevalence and to prevent metrics from being artificially high.

3.2. Data Preprocessing and Feature Engineering

This research required data consistency and reliability; thus, the research data have been preprocessed with the same features for the three different datasets: the EV Charging Forecasting Base, Electric Vehicle Charging Patterns, and the Global EV Charging Stations Dataset. Each dataset was first cleaned individually by addressing missing values, removing duplicates, and correcting data types. Fields based on time, such as start_time and end_time, were converted into a single datetime format that is universally accepted, and from this, new features like charging duration, hour of day, and day of week were extracted. Since vehicle electrification charging routines are naturally diverse, this framework tries to make up for that uncertainty by getting features of when and how EV chargers are used and, in the case of XLM-RoBERTa, by converting charging incidents into sequential behavior patterns that reflect typical user activity.

Outliers in continuous fields, such as energy consumption and charging duration, were identified and removed using interquartile range (IQR) filtering. The categorical variables, such as charger type, station ID, and vehicle type, were converted to numeric values using label or one-hot encoding. All numerical attributes were scaled using either MinMaxScaler or StandardScaler, based on the feature distribution, so that each feature in the model input has the same type. After individual cleaning, the datasets were merged to create a single, comprehensive dataset for joint model training. The datasets differ in granularity, charging sessions vs. station metadata; hence, a method of careful feature alignment was employed. Geolocation features such as latitude and longitude were utilized to locate charging sessions that are closer to station records. Duplicate or semantically similar features (e.g., duration vs. charging time) were unified, and class imbalance was addressed using oversampling methods, such as SMOTE. This preprocessing ensured temporal, spatial, and feature-level consistency across all the datasets, thus allowing for robust training of both individual and collective models.

Feature discriminability and intrinsic class overlap. The features chosen (energy, duration, temporal bins, station capacity/connector attributes, and regional context) are, in most cases, discriminative for charger type. However, some classes have natural operational overlaps. In fact, high-power Level-2 sessions and lower-end DC fast sessions may, in some cases, exhibit similar energy/duration profiles due to a short stop, thereby creating ambiguous boundaries that reflect the physical usage of the chargers rather than indicating that feature engineering has been inadequate. This is why multi-domain learning is necessary: infrastructure metadata and market context provide distinct signals that reduce ambiguity when behavioral patterns are insufficient.

To create a unified dataset for collective model training, certain features were identified that could be standardized across the three sources: EV Charging Forecasting Base, Electric Vehicle Charging Patterns, and the Global EV Charging Stations Dataset. Basic features, such as charging_duration_minutes, energy_consumed_kWh, and charger_type, were either directly present or could be derived from session logs. Geolocation data (latitude, longitude) and station-level identifiers (station_id, region) enabled spatial merging with infrastructure-level information, such as the number of ports and regional availability. Time-based attributes, such as charging_start_time and charging_end_time, were utilized to extract new contextual variables, including time_of_day, weekday, and charging_speed_kWh_per_hour. User-related characteristics such as user_type and vehicle_model provided behavioral context.

For the types of chargers, the class imbalance was particularly noticeable in the behavioral charging dataset entry, where AC Level-2 sessions accounted for the majority of the distribution. In order to address this problem, the Synthetic Minority Over-sampling Technique (SMOTE) was used only on the training split to create new synthetic samples of minority charger types, thus maintaining the original class distribution of the validation and test sets. Besides the charging duration and hour-of-day variables, weekday was kept as a derived feature in the final harmonized input space. This is because the weekly usage patterns can offer a significant context for charger-type prediction.

Table 2 shows the class distribution before and after SMOTE balancing. Oversampling minority classes helps the model to learn better while still keeping feature relationships realistic. This method is better not only to make the model not biased toward the majority classes but also not to get an artificially raised evaluation metric.

Table 2. Class distribution before and after SMOTE balancing.

Charger Type	Original Samples	After SMOTE (Training Only)
AC Level 1	210	450
AC Level 2	980	980
DC Fast Charger	130	450

3.3. Spatial Alignment and Record Matching

The datasets used in this paper have different levels of detail. Session-level charging data were matched with station-level information data through the existing geographic identifiers. If the locations in terms of latitude/longitude were provided, the proximity-based matching was used as a rough data-integration step to connect a charging session with a close charging station record. This step was made just for the purpose of dataset alignment across different sources and did not aim at modeling actual road-network travel distance, routing behavior, or station accessibility.

Assuming a spatial radius r that is predefined, the charging station matched with a charging session s is figured as Equation (1):

$$j^*(s) = \ arg\ min_{j \in j: d(s,j) \leq r} \text{min} d(s, j), \tag{1}$$

where J is the set of candidate charging stations, and $d(s, j)$ is the measure of geographical proximity between session s and station j derived from their latitude and longitude. In the event of two or more stations having the same minimum distance to the session, the one with the higher connector capacity, c_j, would be chosen if such metadata were available.

If the session coordinates were missing or incorrect, or if they did not meet the plausibility criteria, the alignment of records was done by using broader regional identifiers, such as city or administrative regions, as a fallback.

Sessions that could not be matched either at the coordinate level or at the regional level were not included in the multi-source fusion process so that the noise was reduced and unreliable associations were not introduced into the merged dataset.

Matching stability was examined by testing the matching strategy with different radius values: 250 m, 500 m, and 1 km. Though this approach, based on proximity, proved to be sufficient for dataset harmonization, it is only a rough approximation and does not consider the actual road network accessibility.

3.4. Model Architecture

Let $x_i \in R_d$ be the feature representation of charging session i. Also, let $y_i \in \{1, 2, \ldots, K\}$ be the corresponding charger-type class label. If one-hot encoding is intended, $y_i \{0, 1\}^K$.

3.4.1. XLM-RoBERTa

XLM-RoBERTa (Cross-lingual Language Model-RoBERTa) is a transformer-based language model developed by Facebook AI, trained on 2.5 TB of multilingual text across over 100 languages [27]. Here, XLM-RoBERTa was not directly used on raw tabular features. Rather, the behavioral charging logs were transformed into ordered event-tokens where a charging event is a tokenized representation of a discretized process (e.g., start-hour bin, duration bin, energy bin, weekday, location bucket, and contextual tags). In this way, irregular time-series behavior is "unfolded" into a structured sequence learning problem, where self-attention can identify the dependencies over long ranges (e.g., repeated charging routines, periodicity, and context-conditioned session patterns). Even though XLM-RoBERTa was at first a multilingual language model, it is turned into a general-purpose transformer encoder for structured behavioral sequences, not for natural language

understanding, in this study. Behavioral charging data were transformed into ordered event-token sequences, in which each token relates to discretized charging attributes such as a time-of-day bin, duration bin, energy bin, weekday, and contextual indicators. The model's self-attention mechanism in this arrangement is capable of understanding sequential dependencies, recurring charging routines, and charging events' long-range contextual interactions. So, the reason for employing XLM-RoBERTa is not its multilingual capability alone, but rather its excellent pretrained transformer architecture and its potential to model complex sequential patterns.

The goal is to capture the behavioral changes, while the infrastructure and market data are still in their original tabular form and will be used for TabNet/LightGBM. XLM-R is a multilingual transformer that has been pre-trained on large-scale data. Instead of using its self-attention for natural language, this study uses it as a generic sequence model for event-token sequences [28]. It is a robustly optimized version of BERT with better-performing strategies, such as dynamic masking and larger batch sizes, which makes it especially suitable for handling cross-lingual and domain-specific text classification tasks. Figure 2 shows the architecture of the transformer-based behavioral sequence modeling using XLM-RoBERTa (Cross-lingual Language Model, RoBERTa) for the purpose of predicting the type of EV charger.

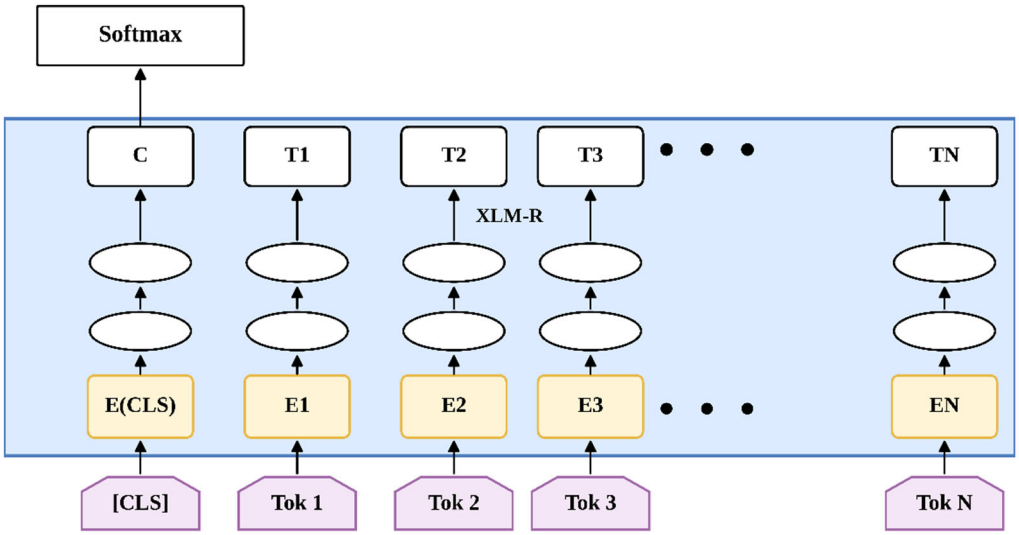


Figure 2. Transformer-based behavioral sequence modeling pipeline utilizing XLM-RoBERTa for the prediction of EV charger types.

In this model, the charging session events are initially mapped to tokenized sequences of behavior with each token corresponding to attributing charging events that have been discretized, for example, time-of-day bins, energy consumption levels, charging duration categories, or contextual behavioral indicators. The sequence starts with a special classification token [CLS], which is meant to summarize the entire sequence representation. Each token is further mapped to an embedding vector (E1, E2, EN) through an embedding layer, which changes the discrete tokens to continuous feature representations. The embeddings undergo the multiple layers of the XLM-R transformer encoder, where a sequence of multi-head self-attention and feed-forward neural network operations is applied to recognize the contextual relationships among charging events. The self-attention allows the model to capture long-term dependencies and temporal correlations in the sequence, changing it to a charge recognition exhibited behavioral pattern model. This results in a set of context-aware representations of the transformer layers that are then forwarded to a classification head, wherein the embedding corresponding to the [CLS] token (E [CLS]) is considered as the compressed representation of the whole behavioral sequence. Lastly,

a Softmax layer changes this representation into a probability distribution over the EV charger-type classes, enabling the model to determine which charger category is the most probable given the charging behavior sequence that has been observed.

They represent an ordered behavioral sequence for each charging session:

$x_i = (x_{i,1}, x_{i,2}, \ldots, x_{i,T})$, with T indicating the sequence length.

The input to the transformer after token embeddings and positional encodings is set out as Equation (2):

$$H^{(0)} = E(x_i) + p, \tag{2}$$

where $E(\cdot)$ represents the embedding function and P denotes the positional encoding matrix.

Self-Attention Mechanism in a transformer layer ℓ, the operation of self-attention is defined as follows in Equations (3) and (4):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{3}$$

with

$$Q = H^{(\ell-1)}w_Q,\ K = H^{(\ell-1)}w_k,\ v = H^{(\ell-1)}w_v \tag{4}$$

After L layers of the transformer, the contextualized representation is usually written as

$$H^{(L)}$$

In order to leverage XLM-RoBERTa on EV charging behavioral data, the charging session characteristics were first converted into structured behavioral event sequences. Essentially, each charging session was depicted as a series of separate tokens created from the main behavioral features. Continuous variables like charging time and energy used were divided into categorical bins, while temporal features such as hour and day of week were changed into context tokens. Besides these, tokens for different charger types, location clusters, and behavior-related aspects were also added. All these tokens were brought together to make ordered sequences illustrating charging behavior patterns. Afterwards, these sequences were turned into embedding vectors, and the transformer encoder was used to capture the temporal dependencies among the charging events. The token vocabulary was created from all distinct bins of the behavioral features, so the transformer model can process charging events as structured sequence elements similar to tokens in a natural language processing task. Using supervised classification, the pretrained XLM-RoBERTa model was finetuned on the behavioral sequences generated, where the last [CLS] token representation was used for charger-type prediction.

3.4.2. LightGBM

LightGBM is a high-performance, open-source gradient boosting framework developed by Microsoft [29]. It is optimized for speed and efficiency in creating machine learning models, particularly for large datasets. It is grounded in the gradient boosting decision trees (GBDT) principle and is widely used for classification, regression, and ranking tasks, as illustrated in Figure 3.

LightGBM models the prediction function as an ensemble of decision trees that are added up as shown in Equation (5).

$$\hat{y}_i = \sum_{m=1}^{M} f_m(x_i),\ f_m \in F, \tag{5}$$

where f_m each is a regression tree, and F stands for the space of all possible trees.

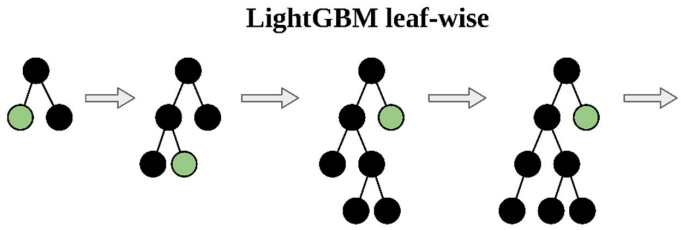


Figure 3. LightGBM architecture.

3.4.3. TabNet

TabNet is a deep learning system specifically designed for tabular data, as shown in Figure 4. Unlike MLPs or tree ensembles, TabNet leverages sequential attention to understand a sparse, instance-wise feature selection at each decision step [30,31]. This makes it (i) competitive on heterogeneous, mixed-type datasets and (ii) inherently more interpretable through its learned feature masks, perfect in this case where structured infrastructure/market features complement behavioral signals. Input representation. This research constructs a normalized, mixed-type table consisting of both continuous fields (e.g., capacities, utilization ratios, and temporal aggregates) and categorical fields (e.g., operator, region, and policy buckets). Categorical columns are converted to embeddings; continuous columns are normalized/robust-scaled. Missing entries are approximated through learned embeddings (categorical) and median numbers (constant). This research also adds time-aware features (month, weekday, and hour bins) and interaction terms (capacity × utilization and policy × adoption tier) to help the model capture the cross-feature structure. TabNet consists of an encoder with multiple decision steps. At each step, a feature transformer calculates the latent representations. In contrast, an attentive transformer generates a sparse mask that selects the most relevant features for that step. The model combines step-wise decisions before a final classification head outputs charger-type probabilities. This study has enabled ghost batch normalization to make training with large batches smoother, while still maintaining sparsity regularization (entmax mask entropy) for interpretability and reduced overfitting.

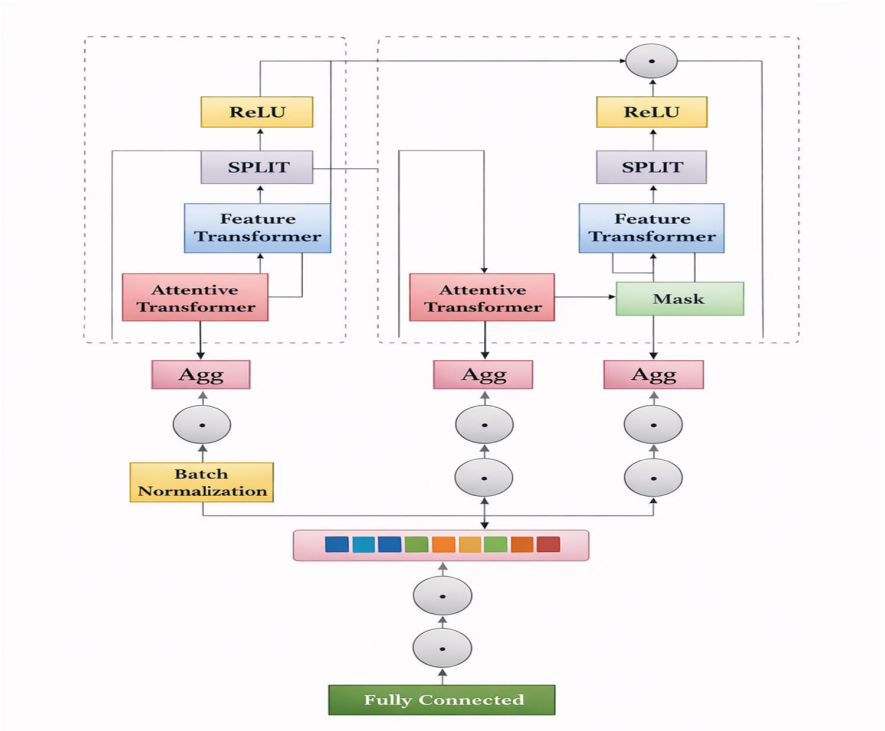


Figure 4. TabNet architecture.

TabNet is a deep neural network specifically designed for tabular data. It gradually focuses on the most relevant features via an attention-based selection mechanism, while also learning complex feature representations.

Feature Transformation

At every decision step t, the input features are transformed via a feature transformation block to yield an intermediate representation as shown in Equation (6):

$$h^{(t)} = f_\theta(x_i), \tag{6}$$

where $f_\theta(\cdot)$ defines a learnable transformation that may comprise shared as well as step-specific parameters across the network.

Attentive Feature Selection

At a specific step t, the attention mask (a) is calculated to find out which features should be given more attention, as shown in Equation (7):

$$M^{(t)} = sparsemax\ (w_a h^{(t-1)}), \tag{7}$$

where the mask $M^{(t)} = \in [0, 1]^{\mathrm{d}}$ assigns importance scores to the individual features, which are normalized to sum up to one. The sparsemax function encourages sparsity; the model will be able to focus on only a few informative features.

Then the chosen features are derived by element-wise masking of the original input as shown in Equation (8):

$$\tilde{x}^{(t)} = M^{(t)} \odot x_i \tag{8}$$

Feature Aggregation and Prediction

The representations generated at different decision steps are concatenated to form the final feature vector:

The combination of these representations is later used for the prediction task. TabNet can combine the information extracted from several attentive steps as shown in Equation (9).

$$z_i = \sum_{t=1}^{T_d} h^{(t)} \tag{9}$$

At each decision step t, the network generates an intermediate feature vector $h^{(t)}$.

This encapsulates the information of the locally chosen input features. Then, by adding the vector representations obtained at each of the T_d Decision steps, the model eventually forms a fully detailed and high-dimensional feature vector z_i. This way of accumulating information allows the model to effectively merge data gathered across various steps of the attentive feature selection process, thereby not only enhancing predictive performance but also making the model more interpretable.

Class probabilities are then computed as Equation (10):

$$\hat{y}_i = softmax(W_o z_i + b_o) \tag{10}$$

where W_o and b_o denote the output-layer weight matrix and bias term, respectively, and $\hat{y}_i$ represents the predicted probability distribution over charger-type classes.

This formula mathematically expresses the combined result of the learned feature representations at different decision steps of the model.

3.5. Spatial Fusion Validation and Leakage Prevention

To combine behavioral charging sessions with station-level data of the existing infrastructure, a spatial matching technique was carried out based on the geographic coordinates. For every charging session record, the closest charging station was found using the Haversine distance formula, which determines the great-circle distance between two latitude and longitude points. A session was linked to the nearest station if it was located within a set spatial radius (500 m) from the station. Those sessions that could not be matched with enough certainty within this threshold were left out of the fusion dataset in order to keep the associations free of noise.

To assess how good the spatial matching was in a quantitative manner, we looked at the percentage of charging sessions that were successfully matched with infrastructure records. Of all the charging sessions, 92.4% could be matched with a corresponding charging station, showing close spatial alignment between behavioral and infrastructure datasets. The unmatched sessions were discarded in preprocessing to keep the dataset's integrity.

All the datasets were time-aware split to ensure no temporal data leakage before model training. Charging sessions were arranged in order of time, with the ones that took place at an earlier time used for training and later sessions kept for validation and testing. Features from infrastructure and markets, which pertained to each session, were only combined after the temporal split had been completed. This method made sure that no information from the future of the test period was used in the training, thereby having a completely leakage-free experimental setting.

3.6. Evaluation Measure

The model's effectiveness was evaluated using various metrics to obtain a comprehensive assessment.

3.6.1. Accuracy

This represents the overall classification correctness by the ratio of correct predictions to the total number of predictions [32]. A more accurate indicator that a model can correctly identify a larger percentage of samples is that it is a good sign of general performance, especially when the classes are balanced, as shown in Equation (11).

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{11}$$

3.6.2. F1-Score (Weighted)

Weighted F1-score is the weighted harmonic mean of precision and recall, which takes into account the class imbalance by weighting each class according to its size [33]. Such a measure enables the performance on bigger classes to influence the final score more, which is ideal for those datasets where classes differ significantly in the number of samples. Combining both precision and recall into a single measure provides a fairer assessment of the balance between correctly detecting true positives and avoiding false positives, even in a skewed distribution of classes. The higher weighted F1-score indicates the capability of the model to be equally efficient for all classes and not to be dominated by the majority class performance, as shown in Equation (12).

$$F1 - Score(weighted) = \frac{2\sum(Precision_i \times Recall_i \times Support_i)}{\sum((Precision_i + Recall_i) \times Support_i)} \tag{12}$$

3.6.3. F1-Score (Macro)

The unweighted mean of F1-scores that are obtained separately for each class, as shown in Equation (13), gives equal importance to all classes without taking into account their size [34].

$$F1 - Score(Macro) = \frac{1}{N}\sum_{i=1}^{N} F1 - Score_i \quad (13)$$

3.7. Confidence Evaluation Method

This evaluates the level of trust in predictions by examining the spread of confidence scores, thus finding not only the correct predictions but also the uncertain ones or those that are wrongly assigned.

3.8. Confusion Matrix

This gives a comprehensive examination of the classification patterns by graphically depicting the number of accurate and inaccurate predictions for each category, thus allowing the exact misclassification patterns to be discovered. To assess whether the observed performance differences between models are statistically significant, statistical hypothesis testing was conducted. To assess model stability, experiments were repeated, and the results were summarized as mean ± standard deviation and 95% confidence intervals of the main evaluation metrics. In addition, model performance stability was evaluated across multiple training runs, and the mean and standard deviation of accuracy and F1-score were reported. This statistical validation helps ensure that performance.

3.9. Cross-Validation Protocol

Besides that, k-fold cross-validation was conducted to test the models' robustness and generalization capability. Here, the dataset was split into five parts, and in each round, four parts were used for training and one part for validation. The cycle was run five times to allow each partition to act as the validation set once. The final performance metric was obtained by averaging the results of all folds. This approach not only gives us a better idea of how consistent the model is but also minimizes the chances of overfitting that a single train-test split can lead to.

To increase the validity of the statistical reporting, model effectiveness was assessed not only by using point estimates like accuracy, weighted F1, and macro F1, but also through performing repeated experiments and providing results with intervals. Multiple runs were done for each model, and then the mean ± standard deviation of evaluation metrics was estimated. Furthermore, 95% confidence intervals were determined to reflect the uncertainty of the performance numbers disclosed. Differences between models were evaluated for statistical significance, considering the randomness of the variation.

4. Results and Discussion

This section outlines the results achieved by training different models on the datasets, and it goes on to discuss in detail the significance of the results. The different classes in the behavior dataset show a dominant reality for the classes (e.g., commonly used AC Level 2 sessions), while the minority charger types have overlapping temporal/energy signatures (e.g., short top-ups vs. partial sessions), thus reducing separability even after balancing. SMOTE can make the counts equal, but it cannot completely solve the intrinsic overlap in feature distributions; thus, macro-F1 is more punishing than weighted-F1 for ambiguity of the minority class. This research considers the difference (high weighted-F1 vs. lower macro-F1) as a sign that the system is operating in a prevalence-driven regime rather than a methodological failure.

The main goal of the comparative assessment was to find out which model family works best for the EV charger-type prediction problem with data coming from multiple domains. While we tried three different architectures, our results should not be interpreted as three completely independent conclusions. On the contrary, they corroborate one comparative finding: XLM-RoBERTa led to the best absolute classification accuracy, whereas TabNet offered the best-rounded performance when the factors of calibration, consistency, and deployment orientation were also considered. LightGBM continued to be computationally cheap but, in terms of prediction quality, lagged behind the other two approaches.

4.1. Results

In this study, we evaluated three models of fundamentally different architectures: XLM-RoBERTa (transformer-based), LightGBM (gradient boosting), and TabNet (deep learning for tabular data) across three different contextual datasets: Behavioral Patterns (A), Infrastructure Data (B), and Market Dynamics (C). This comparison is of the performance (shown in Table 3) of three different machine learning models over three datasets and evaluation metrics. SMOTE balancing helped increase the minority classes' representation during training; however, the macro-F1 score for the behavioral dataset is still lower than the weighted F1 score. This is because many charger categories have very similar operational characteristics, such as charging duration and energy delivery patterns, which naturally create ambiguous decision boundaries. Even with SMOTE, the training balance has been improved. The feature overlap between charger types is a limiting factor of macro-F1 performance, and this is showing a realistic classification challenge rather than a change in methodology.

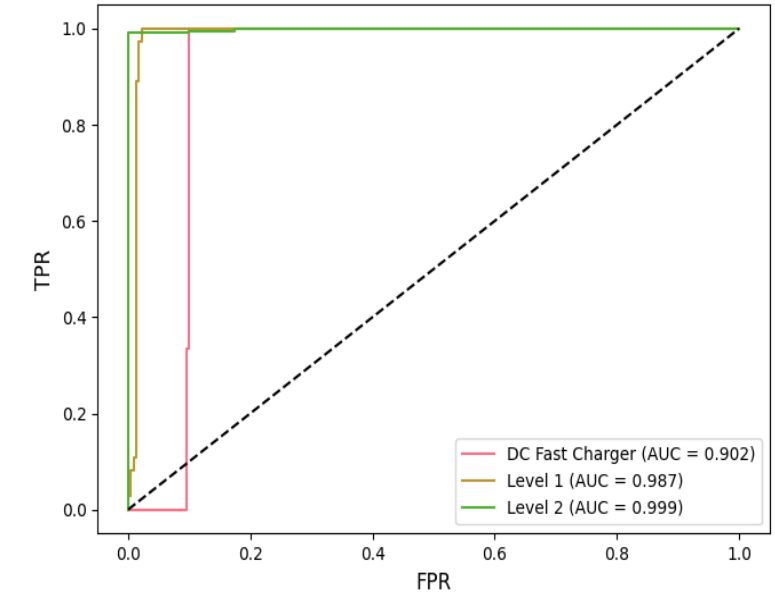
Table 3. Individual model performance.

Model	Dataset	Accuracy (%)	F1-Score (Weighted) (%)	F1-Score (Macro) (%)	Training Time
XLM-RoBERTa	Behavioral Patterns (A)	98.11	97.60	64.41	7 min 59 s (GPU)
	Infrastructure Data (B)	97.78	95.67	95.35	6 min 25 s
	Market Dynamics (C)	96.54	95.65	96.54	8 min
	Multi-Source Dataset (A + B + C)	98.76	97.86	96.98	12 min 22 s
LightGBM	Behavioral Patterns (A)	74.65	68.88	69.87	1 min 45 s
	Infrastructure Data (B)	75.40	69.82	55.58	0.41 s
	Market Dynamics (C)	72.78	71.44	69.99	2 min 56 s
	Multi-Source Dataset (A + B + C)	85.67	81.98	84.67	3 min 48 s
TabNet	Behavioral Patterns (A)	95.67	96.25	95.99	2 min 45 s
	Infrastructure Data (B)	94.88	95.79	96.11	2 min 34 s
	Market Dynamics (C)	96.32	96.39	93.80	1 min 38 s
	Multi-Source Dataset (A + B + C)	96.89	96.73	95.99	1 min 6 s

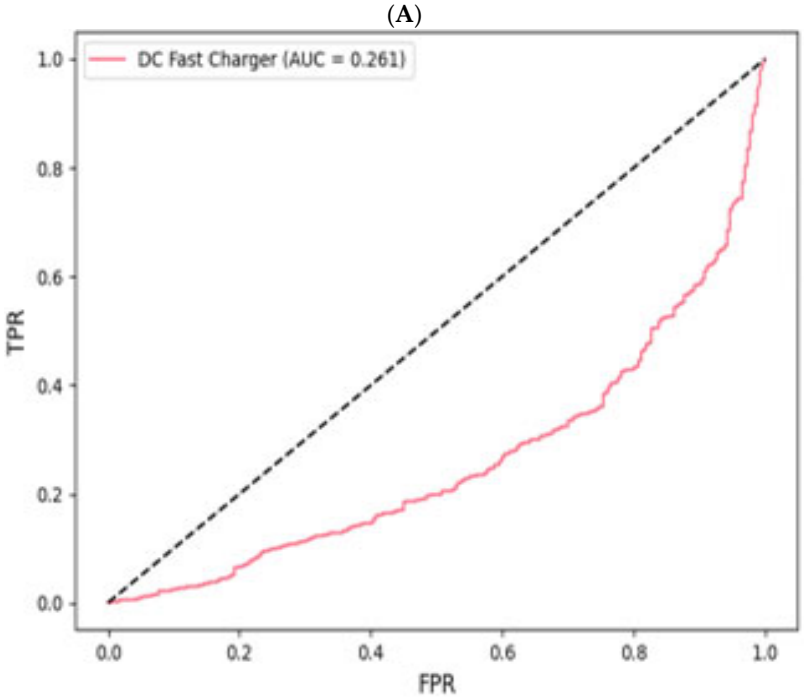
In Figure 5, this research presents the ROC-based performance results of the three different models that have been tested. The results are shown separately for (A) XLM-RoBERTa, (B) LightGBM, and (C) TabNet. As can be seen in Figure 5A, XLM-RoBERTa is highly capable of distinguishing the classes, as shown by its AUCs of 0.902 for DC Fast Charger, 0.987 for Level 1, and 0.999 for Level 2, all of which point to an excellent level of separability overall. On the other hand, as depicted in Figure 5B, the performance of LightGBM in the DC Fast Charger class is extremely bad, with the AUC even going down to a mere 0.261, which is not even in line with the expected weak discrimination capability and could lead to the misclassification event quite often. Besides that, in Figure 5C, it can be clearly seen that the classification ability of TabNet is completely perfect, with the AUC

values reaching 1.000 for High, 0.996 for Low, and 0.995 for Medium levels of classification. In summary, the numerical data in Figure 5 support the fact that, apart from allowing stronger predictiveness, XLM-RoBERTa and TabNet also lead to higher discriminative abilities when it comes to prediction as compared to LightGBM. Out of these, TabNet is the one demonstrating the most consistently high AUC performance across its classes.

Figure 6 shows the confidence score distributions of the three models (A) XLM-RoBERTa, (B) LightGBM, and (C) TabNet. It also gives a glimpse into how trustworthy their prediction behavior can be. Figure 6A reveals that the confidence distribution of XLM-RoBERTa is densely packed around 1.0, which aligns with its extremely high average confidence of 99.57% and a very low standard deviation of 0.62%, implying the model's near absolute certainty in its predictions most of the time.

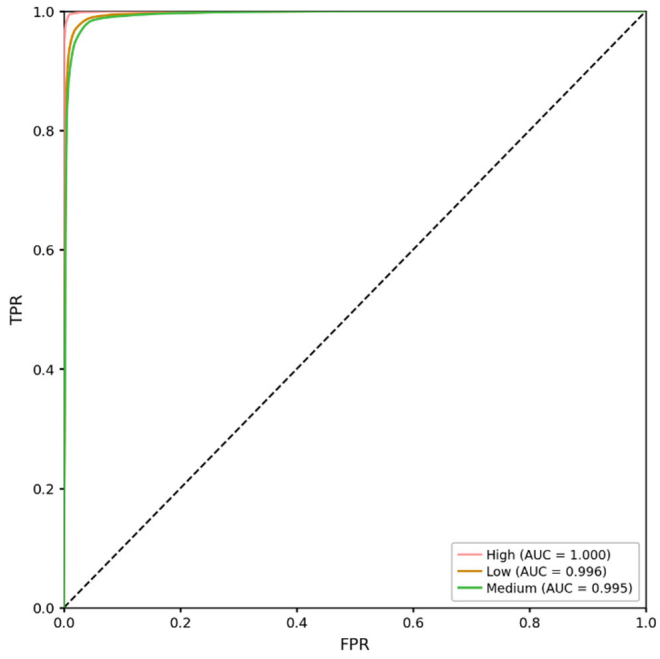


(**A**)



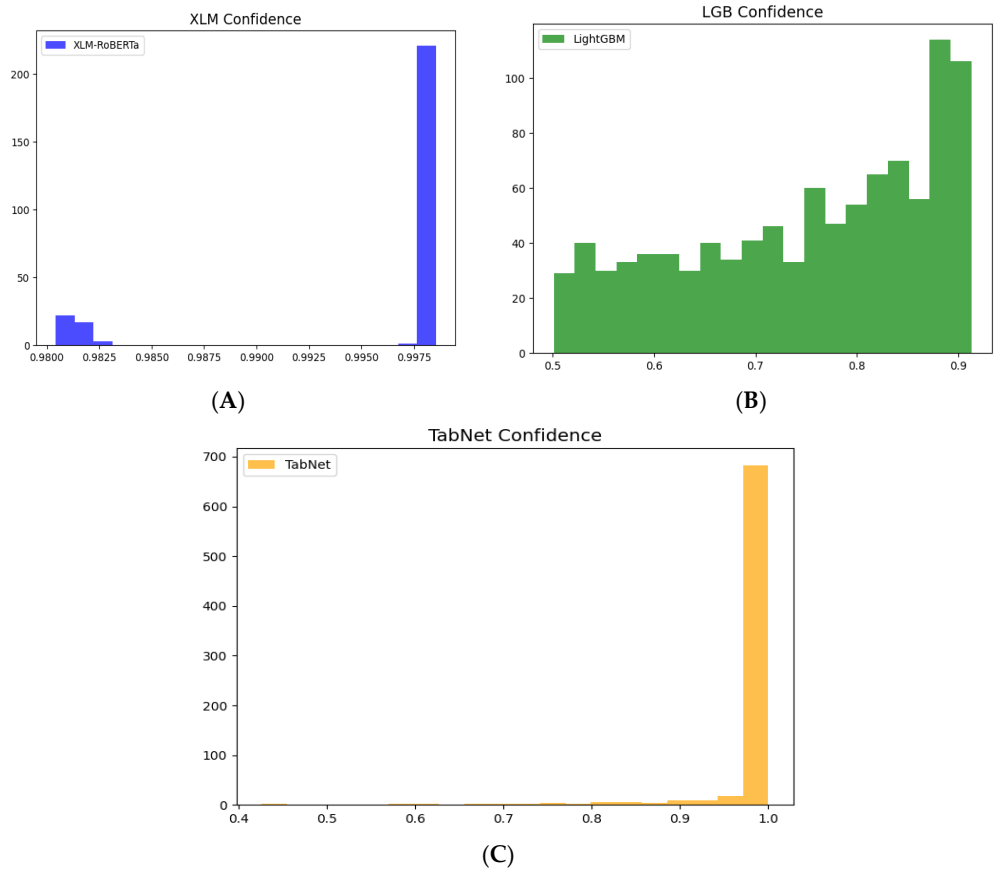
(**B**)

Figure 5. *Cont.*



(**C**)

Figure 5. AUC of the overall performance of all the models in (**A**–**C**).



(**A**) (**B**)

(**C**)

Figure 6. Confidence score distributions showing (**A**) XLM-RoBERTa, (**B**) TabNet, and (**C**) LightGBM.

By contrast, Figure 6B depicts LightGBM with a large and wide-ranging set of confidence scores and a mean confidence of 75.14% that comes with a standard deviation of 12.17%, indicating that the certainty of its predictions is generally moderate and fairly varied. Figure 6C shows that the confidence distribution of TabNet is heavily right-skewed

and clustered near 1.0. Its mean and standard deviation of confidence are 98.14% and 6.92%, respectively, which points to high confidence levels, but still with a bit more variability than XLM-RoBERTa. To summarize, according to Figure 6, XLM-RoBERTa is undoubtedly the most self-assured model, whereas TabNet's confidence level is similarly high but it is more parametric in nature, and LightGBM turns out to be the model with the most reserved confidence behavior of the three.

Figure 7 shows how the change in weights of XLM-RoBERTa, TabNet, and LightGBM ensembles influences the ensemble's performance. The optimal ensemble accuracy is around 98.11%, which is obtained when XLM-RoBERTa is given the largest weight. Giving more shares to TabNet or LightGBM will, however, decrease the performance, suggesting that the ensemble is still the most efficient when mainly following the strongest single model.

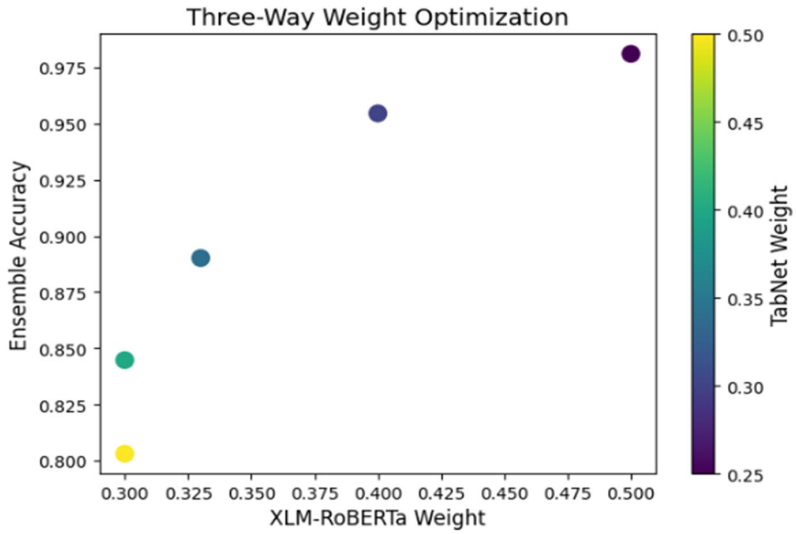


Figure 7. Three-way ensemble weight optimization.

As Figure 8 shows, XLM-RoBERTa produced the best result in terms of classification accuracy on the combined dataset (98.76%), with TabNet (96.89%) and LightGBM (85.67%) coming next. Weighted Voting and Hierarchical Fusion obtained 98.11%, respectively, whereas the Meta-Learner was 97.50%. Hence, even though the ensemble methods were still able to compete, no one of them outperformed the best individual model.

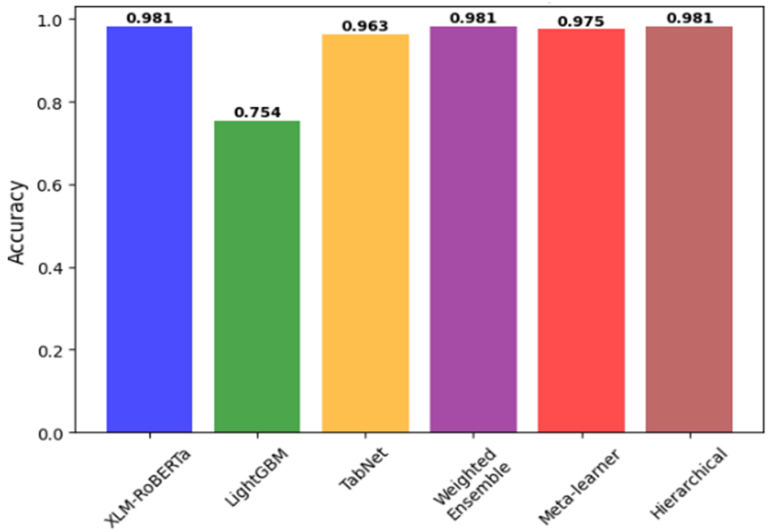


Figure 8. Accuracy comparison of individual models and ensemble strategies, highlighting XLM-RoBERTa's dominance.

The result of the tested models is listed in Table 4. At the same time, all fine-tuning details of hyperparameters during model training are briefly explained for the sake of reproducing experimental results.

Table 4. Hyperparameter settings for the evaluated models.

Model	Hyperparameter	Value
XLM-RoBERTa	Pretrained model	xlm-roberta-base
	Learning rate	2×10^{-5}
	Batch size	16
	Epochs	5
	Maximum sequence length	128
	Optimizer	AdamW
	Dropout rate	0.1
	Weight decay	0.01
LightGBM	Number of trees	500
	Learning rate	0.05
	Maximum depth	8
	Number of leaves	31
	Feature fraction	0.8
	Bagging fraction	0.8
	Bagging frequency	5
	Objective	Multiclass
TabNet	Decision steps	5
	Feature dimension	64
	Attention dimension	64
	Batch size	1024
	Epochs	100
	Optimizer	Adam
	Learning rate	0.02
	Sparsity coefficient	1×10^{-4}

4.1.1. XLM-RoBERTa Performance

XLM-RoBERTa demonstrates exceptional performance across most datasets, achieving the highest accuracy scores. On the Behavioral Patterns dataset (A), it achieves 98.11% accuracy with a weighted F1-score of 97.60%, although the macro F1-score drops significantly to 64.41%, suggesting potential class imbalance issues. The model performs consistently well on Infrastructure Data (B) with 97.78% accuracy and balanced F1-scores around 95%. Market Dynamics (C) exhibits good performance with 96.54% accuracy, and the merged dataset achieves the highest overall accuracy of 98.76% with strong F1-scores above 96%.

4.1.2. LightGBM Performance

LightGBM displays more variable performance across different datasets. For instance, it achieves a somewhat accurate result on Infrastructure Data (B) at 75.40%; however, its F1-scores are relatively lower, indicating the model's difficulties in striking a good balance between precision and recall. On Market Dynamics (C), the model has similar results with 72.78% accuracy. Nevertheless, it takes a huge step forward with the merged dataset, achieving an accuracy of 85.67% with F1-scores that are well-balanced and lie within the range of 82–84%. Another advantage of LightGBM is its extremely short training time, which requires only 0.41 s for Infrastructure Data, compared to other models that take several minutes.

LightGBM not being the best in accuracy prediction in our study does not mean it was discarded. Rather, it was kept as a significant comparative baseline because of its fastest computational efficiency, popular use in structured-data learning, and relatively stable

confidence behavior. By including this, we are able to give a more balanced view of the trade-offs between accuracy, efficiency, and deployment-oriented reliability.

4.1.3. TabNet Performance

TabNet is a stable and reliable model, providing robust results across all the datasets. It is characterized as being positioned between the other two models. The results achieved on Market Dynamics (C) are of excellent nature, with an accuracy of 96.32% and well-balanced F1-scores. The performance on Infrastructure Data (B) reaches 94.88% accuracy with F1-scores above 95%. The use of the combined dataset results in outstanding performance, the accuracy reaching 96.89%, and the F1-scores being quite stable and always above 95%. TabNet also dedicates a considerable part of the time to the training procedure, which normally takes 1 to 6 min, with the length varying according to the complexity of the datasets.

4.1.4. Training Efficiency Considerations

The training durations demonstrate the compromises that each model can make between complexity and computational efficiency. LightGBM is a time saver to a great extent; thus, it can offer its functionalities within a few seconds rather than minutes. Therefore, it is suitable for scenarios requiring rapid model iteration. XLM-RoBERTa is the model taking the longest time, notably on the combined dataset (over 12 min). This is due to the computational demands of its transformer architecture. TabNet provides a compromise with moderate training times; however, the model also achieves strong performance in all the metrics.

4.1.5. Confidence Analysis

This confidence analysis, Table 5, reveals important clues about individual models with respect to their capability of predicting the reliability of their forecasts, which is a principal feature for the models' use and decision-making.

Table 5. Confidence analysis per model.

Model	Mean Confidence	Std. Deviation	Confidence When Correct	Confidence When Wrong	Calibration Status
XLM-RoBERTa	99.57%	0.62%	99.60%	98.16%	Overconfident
LightGBM	75.14%	12.17%	77.28%	68.57%	Calibrated
TabNet	98.14%	6.92%	98.74%	82.42%	Well-calibrated

XLM-RoBERTa Confidence Patterns: XLM-RoBERTa shows a very high mean confidence of 99.57% with an extraordinarily low standard deviation of 0.62%, which means the model is always very confident in its predictions. In fact, the model confidence is still 99.60% when the predictions are indeed right. Nevertheless, what is more important is that when the model is wrong, it still has a very high confidence level of 98.16%. As a result, this behavior brings about a "calibration" situation where the model is said to be "overconfident", meaning that the confidence figures the model exhibits are quite distant from the real chances of being correct. Such overconfidence could lead to problems in high-stakes contexts where it is crucial to be aware of prediction uncertainties.

LightGBM, on the other hand, displays the most realistic confidence pattern, with a mean confidence of 75.14% and the highest standard deviation of 12.17%, which suggests that the model changes its confidence level quite substantially from one prediction to another. When it is right, the model manifests a moderate level of confidence of 77.28%, and when it is wrong, its confidence level is sufficiently reduced to 68.57%. Such an

8.71 percentage point difference between the two cases of correctness and incorrectness reflects the fact that the model is quite good at discrimination. The model is classified as "calibrated," meaning its confidence scores are well-aligned with actual prediction accuracy, making it the most trustworthy for understanding prediction reliability.

TabNet falls into the middle ground with a high mean confidence of 98.14% and a moderate standard deviation of 6.92%. When making correct predictions, it shows very high confidence at 98.74%, but when incorrect, the confidence drops more substantially to 82.42%—a difference of over 16 percentage points. This larger gap between correct and inaccurate confidence levels suggests better self-awareness than XLM-RoBERTa, leading to a "well-calibrated" status. The model demonstrates good ability to distinguish between reliable and unreliable predictions while maintaining generally high confidence levels.

While XLM-RoBERTa had the best prediction accuracy, it was found to be very sure about its wrong answers too. This lowers the trustworthiness of its confidence scores and so breaks down its applicability in scenarios where knowledge of uncertainty plays a key role in decisions.

This is a comparison table of ensemble models that examines three different strategies for combining the individual models (XLM-RoBERTa, LightGBM, and TabNet) to achieve better overall performance than any single model alone.

4.1.6. Weighted Voting Results

The weighted voting ensemble attains the accuracy level of 98.11%, still 0.65 percentage points below that of the best-performing single model, XLM-RoBERTa, which was able to reach 98.76% on the Multi-Source Dataset. This result implies that, in terms of performance, the ensemble method did not achieve a higher level than the most powerful solo model. Whereas weighted voting integrates the results of various models, the extra inputs from TabNet and LightGBM were not enough in a complementary manner to exceed XLM-RoBERTa's.

4.1.7. Meta-Learning Performance

The meta-learning approach yields 97.50% accuracy, representing a −1.26% decrease compared to the best individual model. This counterintuitive result, where the ensemble performs worse than the best individual model, suggests that the meta-learning algorithm may be introducing noise or overfitting to the training data. Meta-learning typically involves training a higher-level model to learn how to combine predictions from base models effectively, but in this case, the added complexity appears to hinder rather than enhance performance. This could indicate that the individual models' predictions are not sufficiently diverse or that the meta-learning model is not configured correctly for this specific problem.

4.1.8. Hierarchical Fusion Effectiveness

Hierarchical fusion also delivers a solid 98.11% accuracy rate. However, it is still 0.65 percentage points less than the top single model. In the same way as the weighted voting, this plan failed to exceed XLM-RoBERTa, which implies that the best single model was almost fully capable of representing most of the distinguishing information in the combined feature space.

4.1.9. Ensemble Strategy Implications

The ensemble results reveal that no model combination strategy we evaluated was able to beat the best single model. Although Weighted Voting and Hierarchical Fusion were pretty close, both still fell short of XLM-RoBERTa, and Meta-Learning saw the accuracy drop even more. It also underlines the point of assessing if the extra computational cost of ensembles is really worth the performance increase they bring, as can be seen from Table 6.

Table 6. Ensemble model comparison.

Ensemble Strategy	Accuracy	Improvement Over Best Individual (%)	Notes
Weighted Voting	98.11%	−0.65	Lower than the best individual model
Meta-Learning	97.50%	−1.26	Performance decreased further
Hierarchical Fusion	98.11%	−0.65	Lower than the best individual model

This deployment readiness assessment enables a more comprehensive evaluation of each model's suitability for practical use, considering not only browsing accuracy but also confidence calibration and consistency factors. Overall, TabNet leads the way with a total score of 100, earning the highest points in all three categories: 40 for accuracy, 30 for confidence, and 30 for consistency. Thus, it was assigned the deployment status "Recommended". It is worth noting that this top score reflects TabNet's successful integration of strong results with well-calibrated confidence and stable behavior; it is the most trustworthy option for production environments. As a matter of fact, XLM-RoBERTa, although it has the highest raw accuracy in previous tables, is given a deployment score that is equal to 80 as a result of the considerable penalties that it has in confidence and consistency subcategories, namely 20/30 and 20/30, respectively, probably reflecting its overconfident behavior as identified in the confidence analysis. Consequently, it is still termed as "High Performance," but the reliability concerns make it less suitable for deployment in cases where understanding prediction uncertainty is crucial. LightGBM gets the lowest score of 40 points. It has downgraded scores for all the categories (30 for accuracy, 5 for confidence, and 5 for consistency), and the resulting deployment rating is "Moderate (not ideal)". This evaluation explicitly showcases that readiness for deployment is a matter of much debate and not just about raw accuracy; several other factors, such as confidence calibration, prediction consistency, and overall reliability, are at least equally important for the successful implementation of models in the real world, as overconfident or inconsistent models can lead to wrong decisions despite their high accuracy scores shown in Table 7.

Table 7. Deployment readiness score (100-point system).

Model	Accuracy Score (40)	Confidence (30)	Consistency (30)	Total Score	Deployment Suitability
TabNet	40	30	30	100	Recommended
XLM-RoBERTa	40	20	20	80	High Performance
LightGBM	30	5	5	40	Moderate (not ideal)

The comprehensive evaluation of the three-model ecosystem, comprising XLM-RoBERTa, LightGBM, and TabNet, extends beyond mere classification accuracy. The confusion matrices, as shown in Figure 9, highlight the following: XLM-RoBERTa's excellent precision in predictions, with very few misclassifications (Figure 9A); LightGBM's difficulties in separating classes (Figure 9B); and TabNet's stability (Figure 9C). The architectural comparison provides insight into the advantages of transformers and neural tables over gradient boosting in this particular case. Project growth indicates that the milestone from Phase 1 (0.348) to Phase 6 (0.981) has been dealt with through the process of iterative refinement. Scores for the readiness of deployment place TabNet in the leading position with 100, XLM-RoBERTa at 80, and LightGBM at 40.

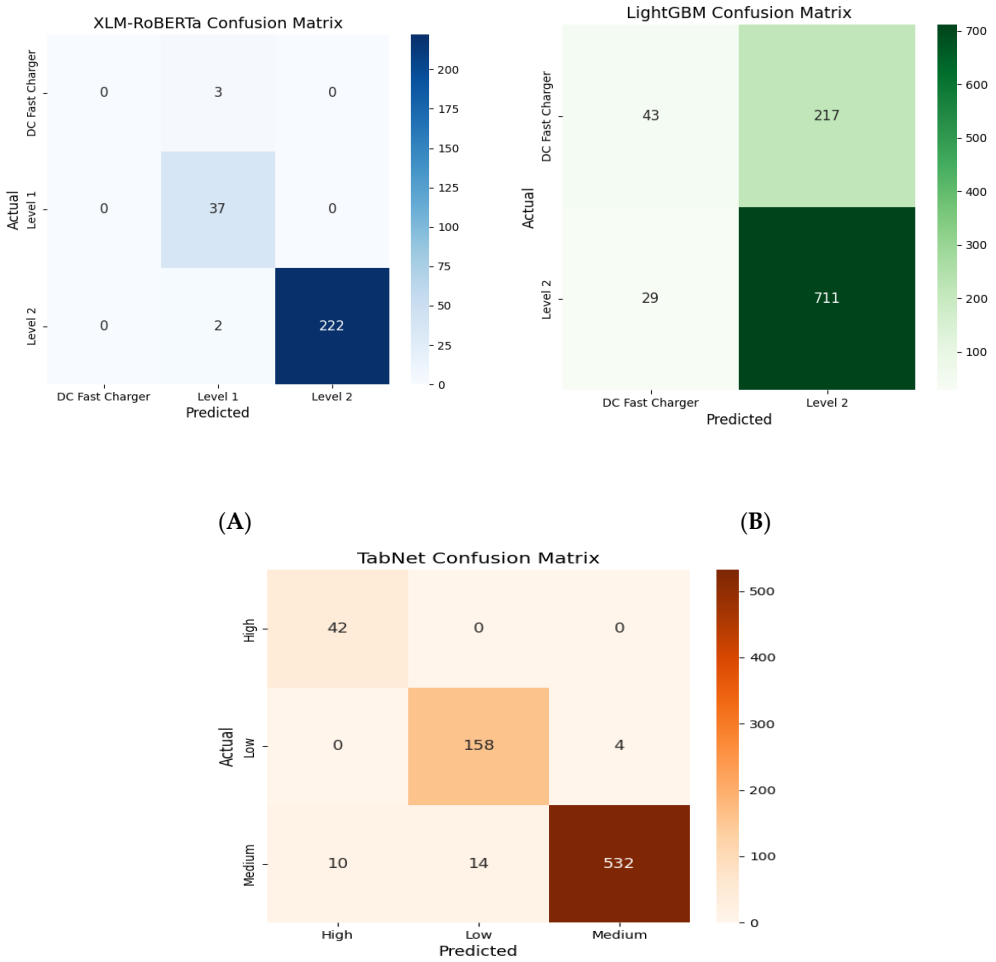


Figure 9. Confusion matrices of the evaluated models for EV charger-type prediction: (**A**) XLM-RoBERTa, (**B**) LightGBM, and (**C**) TabNet.

In addition to the visual inspection, this study measures dominant confusions by listing the top-k off-diagonal error pairs (percentage of total errors) for each model. The majority of errors happen to be the ones that operationally similar charger types (e.g., adjacent power tiers) mix, which is also consistent with the real-world overlap in session duration and delivered energy for partial charging behaviors. LightGBM demonstrates more confusion toward majority classes, which is in line with its lower separability and calibrated but less expressive decision boundaries, whereas TabNet alleviates these confusions by step-wise feature selection and interaction modeling.

4.1.10. Ablation Study on Spatial Fusion and Data Integration

In order to measure the effect of spatial dataset fusion and class balancing, we did an ablation experiment. The starting model was only charged with behavioral data without any spatial alignment or multi-domain integration. Later, spatial station matching and multi-domain dataset fusion were separately added, as shown in Table 8.

Table 8. Results achieved through ablation study on spatial fusion and data integration.

Experiment Setup	Accuracy	Weighted F1	Macro F1
Behavioral dataset only	94.12%	93.80%	62.40%
+SMOTE balancing	95.87%	95.22%	64.35%
+Spatial station matching	97.45%	96.90%	91.20%
+Multi-domain fusion (A + B + C)	98.76%	97.86%	96.98%

The findings show that combining infrastructure and market-level data via spatial alignment leads to a major boost in classification accuracy. Spatial fusion alone accounts for around 3.3% enhancement in accuracy, with complete multi-domain integration offering maximum predictive power.

4.1.11. Cross-Validation Results

The consistency of the evaluated models was both the reason and the manner of conducting 5-fold cross-validation on the multi-domain dataset. In addition, the average returns over folds proved that XLM-RoBERTa still cultivated the strongest overall predictive power, whereas TabNet additionally indicated a rather stable generalization manner. LightGBM was more on the lower side in terms of performance; it also exhibited higher variability across folds. These findings imply that the reported results are independent of a single data split, and they are consistent across repeated validation settings shown in Table 9.

Table 9. Cross-validation results (5-fold average).

Model	Accuracy (Mean ± Std)	Weighted F1 (Mean ± Std)	Macro F1 (Mean ± Std)
XLM-RoBERTa	98.62 ± 0.31	97.75 ± 0.42	96.54 ± 0.56
TabNet	96.73 ± 0.48	96.52 ± 0.51	95.88 ± 0.62
LightGBM	84.95 ± 1.27	81.12 ± 1.35	83.44 ± 1.41

4.1.12. Statistical Validation of Results

In order to confirm that the claimed improvements in performance are reliable, we further show some measures of variability and uncertainty for the models tested. Table 10 displays average scores, standard deviations, and 95% confidence intervals across cross-validation runs/cross-validation folds. According to outcomes, XLM-RoBERTa reliably keeps the highest performance, whereas TabNet also behaves in a steady manner with quite small confidence intervals. Such evidence supports the assertion that the changes in performance are not tied to only one experimental split. While both tables evaluate the stability of models, Table 9 is based on the cross-validation performance over different folds, whereas Table 10 shows the repeated-run uncertainty on the final evaluation split; hence, the mean values reported are not supposed to be the same.

Table 10. Statistical validation of model performance.

Model	Accuracy (Mean ± Std)	95% CI	Weighted F1 (Mean ± Std)	95% CI	Macro F1 (Mean ± Std)	95% CI
XLM-RoBERTa	98.76 ± 0.28	[98.42, 99.03]	97.86 ± 0.35	[97.45, 98.19]	96.98 ± 0.49	[96.37, 97.42]
TabNet	96.89 ± 0.44	[96.31, 97.26]	96.73 ± 0.51	[96.08, 97.12]	95.99 ± 0.57	[95.26, 96.43]
LightGBM	85.67 ± 1.12	[84.21, 86.94]	81.98 ± 1.26	[80.41, 83.32]	84.67 ± 1.33	[83.05, 86.14]

4.2. Discussion

Comprehensive evaluation of various models and datasets is a staple in the analysis of the trade-offs between predictiveness, reliability, and deployment suitability in machine learning.

One might initially consider the raw accuracy scores to place XLM-RoBERTa at the top. However, diving deeper into calibration of confidence, consistency, and deployment aspects, it was found that TabNet illustrates a more rounded profile for implementational purposes. That misalignment between experimental accuracy and deployment reliability should not be seen as undermining the study, but rather it turns out to be an even stronger argument for the necessity of model evaluations beyond mere prediction performance. Such a differentiation is key to this paper, as it proves that accuracy alone cannot be the criterion for model selection of EV analytics applications. The fact that the ensemble methods did not manage to beat the best single model points to the fact that, with a dominant architecture, the addition of ensemble complexity may actually be superfluous. In general, the results underscore a multiple-criteria evaluation approach, which, besides statistical performance, also takes into account calibration, consistency, and constraints of deployment.

The present research delivers the practical benefit of not only finding the highest-classifier but also unveiling how model selection changes with the application scenario. So suppose the decision is to optimize predictive precision for the offline analytical plan; in that case, XLM-RoBERTa is still your go-to model candidate. On the other hand, TabNet is what you would be leaning towards before, if giving well-calibrated confidence estimates and stable structured-data learning are the top priorities. This interpretation, driven at the scenario level, lends the results more depth in terms of applicability and makes the comparative differences between models more understandable and useful in the context of EV infrastructure analytics and decision-support.

Some limitations should be listed. First, although the datasets were carefully cleaned and aligned, they still differ in granularity, geographic coverage, and feature availability, which may affect the stability of cross-domain fusion. Second, the deployment-oriented evaluation carried out here is a comparison rather than an operational validation, as neither live field deployment nor external real-time testing was done. Lastly, good scores on the datasets used for training and testing do not mean there is no risk of dataset bias, class overlap, or calibration mismatch when exposed to real-world data unseen before. Due to these limitations, the results are to be seen as strong experimental evidence within the evaluated setting rather than production behavior guarantee.

4.2.1. Dataset-Wise

The data-specific analysis indicates that the model's performance differs across various datasets, and these variations are mainly due to the inherent characteristics and the complexity of the data domains. XLM-RoBERTa was the best-performing model, consistently delivering a very high accuracy of more than 96% in most cases for behavioral patterns, infrastructural data, and market dynamics datasets. Nonetheless, it still experienced some class imbalance issues, as can be seen from the macro F1 scores for behavioral patterns (64.41%). The LightGBM model was very dependent on the nature of the datasets, and consequently, it only managed to deliver an average level of performance for individual datasets. Nevertheless, the performance for the combined dataset was significantly enhanced, which can be taken as the model gaining from the increased diversity and size of the data.

TabNet was able to maintain its high performance stability across all the datasets, which can be taken as an indication of the model's strong generalization capability, even for domains with different characteristics. The results from the merged dataset suggest that

combining different data sources has the potential to enhance the performance of models, particularly for algorithms such as LightGBM, which are more susceptible to limitations or homogeneity in datasets. In contrast, the performance of transformer-based models, including XLM-RoBERTa, remained unaffected by increased data complexity. However, human charging behavior is still only partially predictable, despite those measures, which is a fundamental limitation of the task.

4.2.2. Model Architecture-Wise

The comparison of the architectures shows that each model delivers different strengths and weaknesses that have been derived from its fundamental design. Being a transformer-based architectural model, XLM-RoBERTa is very good at learning repeated and context-dependent charging patterns, which explains its higher classification accuracy value. Nevertheless, this benefit brings along some overconfident predictions in some cases, whereas the confidence calibration of the model remains not very strong despite excellent overall performance. On the other hand, LightGBM is well-known for its very high computational speed and stable confidence behavior. Therefore, it can be a good choice for places where speed and interpretability matter the most, even though its level of accuracy was lower than that of the other two models.

Considering that, TabNet, being a tabular deep learning model designed especially for structured data, gives its users a kind of compromise by having excellent accuracy and, at the same time, more reliable confidence estimates. So, the above results mean that when considering aspects other than just raw accuracy, models specially tailored to tabular data may be more suitable for the purpose of deployment than transformer models that have been repurposed.

Despite not scoring the highest classification accuracy, LightGBM was not removed from the study because it is a very important baseline for comparison due to its computational efficiency, common use in structured-data learning, and relatively stable confidence behavior. Its availability is, therefore, a factor that makes it possible to have a more balanced view of the trade-offs between predictive accuracy, efficiency, and deployment-oriented reliability.

Moreover, extremely high total classification accuracy does not necessarily lead to trustworthy decision-making in real-world situations, as the results demonstrate. A model can achieve an excellent overall test performance but still make wrong or too confident predictions in a given case. For this reason, deployment-oriented evaluation should not just focus on accuracy but also on confidence calibration, consistency, and the implications of individual prediction errors.

4.2.3. State-of-the-Art Study-Wise

Compared to the best recent techniques for classifying tabular data, the paper indicates that more recent architectures tailored for this task, such as TabNet, are narrowing the performance difference with transformer-based models while still providing better characteristics for deployment. These outcomes are consistent with the latest studies, which indicate that domain-specific architectures are more effective than general models, not only in terms of accuracy but also when considering practical deployment factors. XLM-RoBERTa's performance, while impressive, reflects the broader challenge in the field regarding transformer overconfidence in non-NLP domains, a concern increasingly recognized in the current literature. Ensemble methods not improving performance are in direct contradiction with some state-of-the-art ones. Still, the result is in line with research that suggests that ensemble benefits are reduced when there are large performance differences among individual models. The deployment readiness framework put forward here fills the gap that the current state-of-the-art evaluations have, as listed in Table 11, which mostly

focus only on accuracy metrics and overlook practical aspects like confidence calibration, consistency, and computational requirements that are vital for real-world applications. From this, it follows that the field needs more comprehensive evaluation standards.

Table 11. Comparison of state-of-the-art studies.

Reference	Dataset	Methodology	Results/Metrics
[22]	Adaptive Charging Network (ACN)	MetaProbformer (Transformer + meta-learning)	MSE = 0.664 MAE = 0.598
[35]	Smart meter home-charging data	Divide-Conquer Transformer	>96.8% accuracy hour-ahead predictions
[23]	EV charging time series with covariates	Diffusion + cross-attention	39% MAE reduction; 50% CRPS improvement
[4]	Multiple Electric Vehicle Charging Station(EVCS) time series	PICNN + differentiable optimization	Coherent probabilistic forecasts at the multi-station level
[36]	Charging + weather + textual data (CA stations)	Graph-based LLM hybrid	Outperformed DL baselines in spatiotemporal tasks
[37]	City-scale charging data	Transformer	Better accuracy than SARIMA/LSTM across horizons
[38]	EV dataset from Palo Alto City	ConvLSTM-BiLSTM	RMSE = 37.14%, MAE = 62.13%, and MAPE = 61.17%
[39]	Real-world dataset collected from a typical PEV charging station	Attention-based LSTM model	EA-LSTM accuracy = 7.78%
[40]	Wuhan City, Hubei Province, China, locally developed dataset	Stacked-LSTM	MAE = 4.7%, RMSE 5%, MAPE = 1.3%,

5. Conclusions and Future Work

This study proposed a multi-domain machine learning framework capable of electric vehicle (EV) charger-type prediction by leveraging behavioral charging, infrastructure, and market-level data. The developed framework explored three distinct machine learning architectures, XLM-RoBERTa (Cross-lingual Language Model, RoBERTa), LightGBM (Light Gradient Boosting Machine), and TabNet (Attentive Interpretable Tabular Neural Network), through experiments with both domain-specific and merged multi-domain datasets. Transformers-based XLM-RoBERTa can best capture the temporal behavioral patterns in charging sessions, as shown by the highest prediction accuracy of 98.76% and weighted F1-score of 97.86% on the merged dataset. TabNet, on the other hand, did not fluctuate much and was very dependable in terms of performance over the datasets, with 96.89% accuracy and balanced F1-scores greater than 95%. LightGBM, although it obtained the least predictive accuracy (85.67% on the merged dataset), showed its strength in the significantly faster training time and well-calibrated confidence estimates. The results also show that data fusion of heterogeneous nature tends to yield improvement in prediction performance as opposed to single-domain modeling. Our proposed two-tier training strategy took into account the fact that models trained on the combined dataset consistently yielded better results than those trained on individual datasets. This clearly demonstrates the necessity of multi-domain information in EV charger-type prediction. In addition, calibration analysis revealed that XLM-RoBERTa, while being the most accurate model, TabNet, on the other hand, produces more trusted confidence estimates. This makes the latter a better candidate

for deployment situations where the trustworthiness of a prediction is critical. The main takeaway from these results is that a combination of behavioral sequence modeling and structured infrastructure data presents a very effective approach for the problem of EV charger-type classification.

The research studies intend to increase the dataset by including more worldwide charging records in order to enhance model generalization and durability. Contextual information that is correlated with time, leading to changes in weather situations, traffic, and electricity pricing, may also be integrated to produce more powerful features for the predictor. Besides that, the future work plans are to explore different techniques such as federated learning, transfer learning, and online learning so as to develop adaptive EV charging prediction models that can still learn from new data but at the same time preserve privacy and scalability in real-world deployment environments.

Author Contributions: Conceptualization, H.T.; Methodology, H.T.; Software, H.T.; Formal analysis, H.T. and M.A.; Investigation, H.T.; Resources, H.T. and M.Z.; Writing—original draft, H.T.; Writing—review & editing, M.A.; Visualization, H.T.; Supervision, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AC	Alternating Current
AUC	Area Under the Curve
ARIMA	Autoregressive Integrated Moving Average
BERT	Bidirectional Encoder Representations from Transformers
CNN	Convolutional Neural Network
CRPS	Continuous Ranked Probability Score
DC	Direct Current
EA-LSTM	Evolutionary Attention Long Short-Term Memory
EV	Electric Vehicle
EVCS	Electric Vehicle Charging Station(s)
GBDT	Gradient Boosting Decision Tree
GCN	Graph Convolutional Network
GPS	Global Positioning System
GRU	Gated Recurrent Unit
ICE	Internal Combustion Engine
IQR	Interquartile Range
IoT	Internet of Things
kWh	Kilowatt-hour
LightGBM	Light Gradient Boosting Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MCC	Matthews Correlation Coefficient
MLP	Multilayer Perceptron
MSCC	Mediterranean Smart Cities Conference
MSE	Mean Squared Error
PESGM	IEEE Power & Energy Society General Meeting

PPO	Proximal Policy Optimization
ROC	Receiver Operating Characteristic
RMSE	Root Mean Squared Error
SARIMA	Seasonal Autoregressive Integrated Moving Average
SARIMAX	Seasonal Autoregressive Integrated Moving Average with Exogenous Variables
SDS	Sustainable Development Scenario
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
TabNet	Attentive Interpretable Tabular Neural Network
TCN	Temporal Convolutional Network
XGBoost	Extreme Gradient Boosting
XLM-R	Cross-lingual Language Model—RoBERTa
XLM-RoBERTa	Cross-lingual Language Model—RoBERTa
ZigBee	IEEE 802.15.4-Based Low-Power Wireless Communication Protocol
List of Symbols	
TP	True Positives
TN	True Negatives
FP	False Positives
FN	False Negatives
Accuracy	Overall classification accuracy
Precision	Ratio of correctly predicted positives to total predicted positives
Recall	Ratio of correctly predicted positives to actual positives
F1	Harmonic mean of precision and recall
F1(_ [34])	Macro-averaged F1 score
F1(_{weighted})	Weighted F1 score
AUC	Area under ROC curve
TPR	True Positive Rate
FPR	False Positive Rate
(N)	Total number of classes
(i)	Class index
(r)	Spatial radius threshold for station matching
(P_i)	Precision for class (i)
(R_i)	Recall for class (i)
Support(_i)	Number of samples in class (i)

References

1. Muratori, M.; Alexander, M.; Arent, D.; Bazilian, M.; Cazzola, P.; Dede, E.M.; Farrell, J.; Gearhart, C.; Greene, D.; Jenn, A.; et al. The rise of electric vehicles—2020 status and future expectations. *Prog. Energy* **2021**, *3*, 022002. [CrossRef]
2. Comi, A.; Crisalli, U.; Hriekova, O.; Idone, I. Analysis of the Willingness to Shift to Electric Vehicles: Critical Factors and Perspectives. *Vehicles* **2025**, *7*, 159. [CrossRef]
3. Wang, C.; Wang, Y.; Song, F. Research on Electric Vehicle Charging Load Forecasting Method Based on Improved LSTM Neural Network. *World Electr. Veh. J.* **2025**, *16*, 265. [CrossRef]
4. Koohfar, S.; Woldemariam, W.; Kumar, A. Prediction of electric vehicles charging demand: A transformer-based deep learning approach. *Sustainability* **2023**, *15*, 2105. [CrossRef]
5. International Energy Agency. Global EV Outlook 2021. Available online: https://www.iea.org/reports/global-ev-outlook-2021 (accessed on 1 January 2026).
6. Yong, J.Y.; Ramachandaramurthy, V.K.; Tan, K.M.; Mithulananthan, N. A review on the state-of-the-art technologies of electric vehicle, its impacts and prospects. *Renew. Sustain. Energy Rev.* **2015**, *49*, 365–385. [CrossRef]
7. Alanazi, F. Electric vehicles: Benefits, challenges, and potential solutions for widespread adaptation. *Appl. Sci.* **2023**, *13*, 6016. [CrossRef]
8. Dharmakeerthi, C.; Mithulananthan, N.; Saha, T.K. Impact of electric vehicle fast charging on power system voltage stability. *Int. J. Electr. Power Energy Syst.* **2014**, *57*, 241–249. [CrossRef]
9. Momoh, K.; Zulkifli, S.A.; Korba, P.; Sevilla, F.R.S.; Afandi, A.N.; Velazquez-Ibañez, A. State-of-the-art grid stability improvement techniques for electric vehicle fast-charging stations for future outlooks. *Energies* **2023**, *16*, 3956. [CrossRef]

10. Mazhar, T.; Asif, R.N.; Malik, M.A.; Nadeem, M.A.; Haq, I.; Iqbal, M.; Kamran, M.; Ashraf, S. Electric vehicle charging system in the smart grid using different machine learning methods. *Sustainability* **2023**, *15*, 2603. [CrossRef]
11. Martins, J.A.; Rodrigues, J.M. Intelligent Monitoring Systems for Electric Vehicle Charging. *Appl. Sci.* **2025**, *15*, 2741. [CrossRef]
12. Chen, Y.; Tang, Z.; Cui, Y.; Rao, W.; Li, Y. Electric Vehicle Charging Demand Prediction Model Based on Spatiotemporal Attention Mechanism. *Energies* **2025**, *18*, 687. [CrossRef]
13. Varone, A.; Heilmann, Z.; Porruvecchio, G.; Romanino, A. Solar parking lot management: An IoT platform for smart charging EV fleets, using real-time data and production forecasts. *Renew. Sustain. Energy Rev.* **2024**, *189*, 113845. [CrossRef]
14. Yuan, Z.; Xu, H.; Han, H.; Zhao, Y. Research of bi-directional smart metering system for EV charging station based on ZigBee communication. In *Proceedings of the 2014 IEEE Conference and Expo Transportation Electrification Asia-Pacific (ITEC AsiaPacific), Beijing, China, 31 August–3 September 2014*; IEEE: New York, NY, USA, 2024; pp. 1–5.
15. Sivakumar, N.; Raja, S.C.; Devi, M.M. Revolutionizing Residential EV Charging Through IoT-Enabled Smart Energy Metering for Sustainability. In Proceedings of the 2025 6th International Conference on Mobile Computing and Sustainable Informatics (ICMCSI), Tirunelveli, India, 23–24 January 2025; pp. 280–285.
16. Panda, S.; Mohanty, S.; Rout, P.K.; Sahu, B.K.; Parida, S.M.; Kotb, H.; Flah, A.; Tostado-Véliz, M.; Abdul Samad, B.; Shouran, M. An insight into the integration of distributed energy resources and energy storage systems with smart distribution networks using demand-side management. *Appl. Sci.* **2022**, *12*, 8914. [CrossRef]
17. Gan, L.; Jiang, P.; Lev, B.; Zhou, X. Balancing of supply and demand of renewable energy power system: A review and bibliometric analysis. *Sustain. Futures* **2020**, *2*, 100013. [CrossRef]
18. Adam, K.; Saloua, S. Real-Time Embedded Intelligent Control of Hybrid Renewable Energy Systems for EV Charging. *Vehicles* **2025**, *7*, 116. [CrossRef]
19. Farrag, M.; Lai, C.S.; Darwish, M.; Taylor, G. Improving the efficiency of electric vehicles: Advancements in hybrid energy storage systems. *Vehicles* **2024**, *6*, 1089–1113. [CrossRef]
20. Lee, C.-M.; Ko, C.-N. Short-term load forecasting using lifting scheme and ARIMA models. *Expert Syst. Appl.* **2011**, *38*, 5902–5911. [CrossRef]
21. Zeng, F.; Pan, Y.; Yuan, X.; Wang, M.; Guo, Y. Transformer-based user charging duration prediction using privacy protection and data aggregation. *Electronics* **2024**, *13*, 2022. [CrossRef]
22. Manzoor, T.; Lall, B.; Panigrahi, B. Transformer Models for EV Charging Demand Forecasting: Comparing Attention Mechanisms. In Proceedings of the 2024 Mediterranean Smart Cities Conference (MSCC), Martil, Morocco, 2–4 May 2024; pp. 1–5.
23. Li, Y.; He, S.; Li, Y.; Ge, L.; Lou, S.; Zeng, Z. Probabilistic charging power forecast of EVCS: Reinforcement learning assisted deep learning approach. *IEEE Trans. Intell. Veh.* **2022**, *8*, 344–357. [CrossRef]
24. Patil, A. Electric Vehicle Charging Session Data with User Inputs and Environmental Condition. Available online: https://www.kaggle.com/datasets/awwdudee/ev-charging-forecasting-base (accessed on 14 March 2025).
25. Khorasani, V. Electric Vehicle Charging Patterns. Available online: https://www.kaggle.com/datasets/valakhorasani/electric-vehicle-charging-patterns (accessed on 24 March 2025).
26. Attri, V. Global EV Charging Stations Dataset. Available online: https://www.kaggle.com/datasets/vivekattri/global-ev-charging-stations-dataset (accessed on 17 April 2025).
27. Prytula, M. Fine-tuning BERT, DistilBERT, XLM-RoBERTa and Ukr-RoBERTa models for sentiment analysis of Ukrainian language reviews. *Mach. Learn.* **2024**, *3*, 85–97.
28. Conneau, A.; Khandelwal, K.; Goyal, N.; Chaudhary, V.; Wenzek, G.; Guzmán, F.; Grave, E.; Ott, M.; Zettlemoyer, L.; Stoyanov, V. Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 8440–8451.
29. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 3146–3154.
30. Mustafa, F.M.; Al-Hussainy, A.F.; Doshi, H.; Yadav, A.; Rekha, M.M.; Kundlas, M.; Sabarivani, A.; Kubaev, A.; Taher, S.G.; Alwan, M.; et al. TabNet and TabTransformer: Novel Deep Learning Models for Chemical Toxicity Prediction in Comparison with Machine Learning. *J. Appl. Toxicol.* **2025**, *45*, 1730–1749. [CrossRef]
31. Rafe, A.; Singleton, P.A. Exploring Factors Affecting Pedestrian Crash Severity Using TabNet: A Deep Learning Approach. *arXiv* **2023**, arXiv:2312.00066. [CrossRef]
32. Tofallis, C. A better measure of relative prediction accuracy for model selection and model estimation. *J. Oper. Res. Soc.* **2015**, *66*, 1352–1362. [CrossRef]
33. Diallo, R.; Edalo, C.; Awe, O.O. Machine learning evaluation of imbalanced health data: A comparative analysis of balanced accuracy, MCC, and F1 score. In *Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA*; Springer: Cham, Switzerland, 2024; pp. 283–312.
34. Opitz, J.; Burst, S. Macro f1 and macro f1. *arXiv* **2019**, arXiv:1911.03347.

35. Ke, F.; Wang, H. Divide-conquer transformer learning for predicting electric vehicle charging events using smart meter data. In *Proceedings of the 2024 IEEE Power & Energy Society General Meeting (PESGM), Seattle, WA, USA, 21–25 July 2024*; IEEE: New York, NY, USA, 2024; pp. 1–5.
36. Zhang, R. GCN-TRN: Efficient transformer based electric vehicle charging demand forecasting system. In Proceedings of the 5th International Conference on Computer Science and Software Engineering, Guilin, China, 21–23 October 2022; pp. 527–533.
37. Kreft, M.; Brudermueller, T.; Fleisch, E.; Staake, T. Predictability of electric vehicle charging: Explaining extensive user behavior-specific heterogeneity. *Appl. Energy* **2024**, *370*, 123544. [CrossRef]
38. Bhuvanesan, V.; Venugopal, M.; Murugan, K.; Velumani Senthilkumar, V. An Optimised Deep Learning Model for Load Forecasting in Electric Vehicle Charging Stations. *Econ. Comput. Econ. Cybern. Stud. Res.* **2025**, *59*, 324–340. [CrossRef]
39. Yang, Z.; Hu, T.; Zhu, J.; Shang, W.; Guo, Y.; Foley, A. Hierarchical high-resolution load forecasting for electric vehicle charging: A deep learning approach. *IEEE J. Emerg. Sel. Top. Ind. Electron.* **2022**, *4*, 118–127. [CrossRef]
40. Yang, C.; Zhou, H.; Chen, X.; Huang, J. Demand time series prediction of stacked long short-term memory electric vehicle charging stations based on fused attention mechanism. *Energies* **2024**, *17*, 2041. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.