VILNIUS UNIVERSITY
FACULTY OF MATHEMATICS AND INFORMATICS
INFORMATICS STUDY PROGRAMME

Master’s Thesis

Optimizing Electric Vehicle Charging Infrastructure Using Machine Learning Algorithms

Elektromobilių įkrovimo infrastruktūros optimizavimas naudojant mašininio mokymosi algoritmus

Usman Afzal

Supervisor	:	{Dr. Marco Marcozzi}
Faculty		Mathematics and Informatics
Program		Master’s in informatics
Place and year		Vilnius, Lithuania 2026

Acknowledgements

First and foremost, I would like to thank my suffervisor Dr. Marco Marcozzi for his valuable guidance in every step of the t in completing this this without his help his feedback s played an importance role in shaping this thesis from conception to completion.

Secondly, I would like to acknowledge a grateful effort by Vilnius municipality for making these data set publicly available including population and traffic congestion data. Similarly other sources as well including OpenStreetMap and overpass turbo API.

Lastly, I would like to thank Vilnius university for providing me with a wonderful studying environment. I am gratefully

Summery

The global transition towards electric vehicles requires an optimal solution for charging infrastructure planning. By 2030, the global stock of electric vehicles will rise to 250 million; however, the existing charging station placement method relies on heuristic approaches that fail to incorporate clustering, demand forecasting, site classification, and optimization in a signal framework. This thesis develops a multi-stage machine learning framework for optimal electric vehicle charging station placement in Vilnius, Lithuania.

The proposed framework incorporates seniūnija-level population data (636,023 residents, 21 districts), traffic congestion events, and points of interest from OpenStreetMap. Firstly, DBSCAN clustering with convex hull geometry for candidate zone identification, second, an ensemble of Ridge Regression and random forest for demand prediction, third, a support vector machine used for site suitability classification on revealed preference, lastly, a Genetic algorithm is used for optimal station site selection respecting distance constraints.

The framework identified 201 candidate zones. An ensemble achieved an R^2 of 0.96. The support vector machine achieved an accuracy of 87%. The generic algorithm selected 20 charging station locations with a coverage of 95%, covering 20 districts out of 21, with the fitness of o.4387 compared to the greedy baseline.

This research provides a validated data-driven framework for electric vehicle infrastructure planning, providing a comprehensive guideline for urban planners for the fair distribution of charging infrastructure deployment.

Keywords: Electric vehicle charging station, Machine learning for EV planning, DBSCAN, Ensemble, Clustering, Support vector machine, Genetic algorithm

SANTRAUKA

Pasaulinis perėjimas prie elektromobilių reikalauja optimalaus įkrovimo infrastruktūros planavimo sprendimo. Iki 2030 metų pasaulinis elektromobilių skaičius išaugs iki 250 milijonų; tačiau esami įkrovimo stočių išdėstymo metodai remiasi euristiniais metodais, kurie neapima klasterizavimo, paklausos prognozavimo, vietų klasifikavimo ir optimizavimo viename sistemoje. Šiame darbe sukurtas kelių etapų mašininio mokymosi pagrįstas elektromobilių įkrovimo stočių išdėstymo optimizavimo metodas, pritaikytas Vilniaus miestui.

Siūlomą metodą sudaro seniūnijų lygmens gyventojų duomenys (636 023 gyventojai, 21 seniūnija), eismo spūsčių įvykiai ir interesų taškai iš „OpenStreetMap“. Pirma, DBSCAN klasterizavimas su išgaubto korpuso geometrija kandidatės zonų identifikavimui; antra, Ridžo regresijos ir atsitiktinių miškų ansamblis paklausos prognozavimui; trečia, atraminių vektorių mašina, naudojama vietų tinkamumo klasifikavimui pagal atskleistas preferencijas; galiausiai, genetinis algoritmas naudojamas optimalių stočių vietų parinkimui, atsižvelgiant į atstumo apribojimus.

Metodas identifikavo 201 kandidatę zoną. Ansamblis pasiekė R^2 koeficientą 0,96. Atraminių vektorių mašina pasiekė 87% tikslumą. Genetinis algoritmas atrinko 20 įkrovimo stočių vietų su 95% aprėptimi, apimančia 20 iš 21 seniūnijos, pasiekdamas tinkamumo balą 0,4387, palyginti su godžiuoju pradiniu modeliu.

Šis tyrimas pateikia patvirtintą duomenimis pagrįstą elektromobilių infrastruktūros planavimo sistemą, suteikiančią išsamias gaires miestų planuotojams, kaip teisingai paskirstyti įkrovimo infrastruktūros diegimą.

Raktiniai žodžiai: elektromobilių įkrovimo stotelė, mašininis mokymasis elektromobilių planavimui, DBSCAN, ansamblis, klasterizavimas, atraminių vektorių mašina, genetinis algoritmas

1. Table of Contents

INTRODUCTION

The transportation industry is one of the main sectors contributing to Carbon Dioxide emissions, which accounts for 23.5% of total CO_2 emissions (IEA, 2023). As the global stock of electric vehicles (EV) will rise to over 250 million by 2030, this means that the necessary charging infrastructure needs to be developed intensively. The adoption of electric vehicles offers a promising solution in reducing carbon emissions. The existing charging infrastructure that accommodates EV-based transportation is deficient.

One of the most important factors that facilitates the adoption of EVs is the installation of infrastructure. Cities and transport organizations are under increasing pressure to develop scalable, equitable, and accessible infrastructure due to the expansion of EV markets.

Practically, the electric vehicle charging station (EVCS) infrastructure includes environmental impact assessment, traffic flow optimization, user access, space availability, and electric grid integration (Liu13a). The land required for infrastructure for the spatial plan is difficult to acquire due to regulatory obstacles, and political interests that may not align with current user demand or trends in future urban expansion [Fra11a]. These restrictions resulted in anomalies between the demand for charging and the supply of infrastructure, which reduces network effectiveness and user satisfaction.

The use of infrastructure is greatly affected by human behavior in addition to physical infrastructure and grid boundaries. Where and when users choose to charge their cars, it depends on many factors, including residential habits, domestic types, income levels and access to residential parking. However, existing models often fail to sufficiently include these factors. Ignoring socio-economic factors may cause congestion in high-income districts, due to which EV ownership or low-income areas may get concentrated [LiL16a]. Such multidimensional data can be incorporated into improved machine learning (ML) models, which allows for a more efficient and effective plan. The regulatory framework also has a significant impact on the EVCS plan. The infrastructure design of the charging station is mainly influenced by different factors, such as energy, environmental, and subsidy programs. While some locations encourage gas station retrofitting, and other recommend incorporating stations with existing solar energy.

In general, the EVCS infrastructure placement problem includes different sectors of power engineering, spatial analysis, environmental science, public policy, and data science.

As the growth of EV adoption is rising, there is a subsequent need to develop or redesign new EVCS infrastructure. To address this issue, this study proposed a data-driven methodology considering social and economic contributions to modern urban planning.

Existing EVCS infrastructure planning approaches widely depend on simplified heuristics or static demand estimation, which is inefficient due to the multi-dimensional nature of urban activity.

To address this gap, the research work “Optimizing Electric Vehicle Charging Infrastructure Using Machine Learning Algorithms” is proposed.

Relevance and Novelty: The adoption of electric vehicles provides a promising solution to reduce the problem of carbon emissions globally. Furthermore, the basic infrastructure required to facilitate EV-based mobility is deficient in the urban environment. Most of the existing research depends on simplified distance-based heuristics or static demand estimation and is unable to consider the multi-dimensional urban activity behavior, leading to poor planning of charging infrastructure. To address this multidimensional issue, this study proposes a unified multi-stage framework that uses real-world geospatial data.

As different traditional models consider site selection as a single predictive task, the proposed methodology introduces a multistage pipeline using different datasets, which includes heterogeneous urban data, real traffic congestion events, and Point of Interest POI to identify urban demand at the micro-spatial level.

This study uses a s hybrid approach instead of simply depending on poi counts. Demand estimation is identified by using dual-tier DBSCSN clustering and convex hull-based footprint engineering to identify high-activity areas in the city. Furthermore, an ensemble predictive engine is used for further refinement by providing robust demand projection while mitigating the risk of overfitting.

The distinctive feature of this strategy is the integration of SVM as an empirical feasibility filter, trained on revealed preferences of existing charging station operators.

The model learns the spatial characteristics from the already installed charging station locations by different operators. Which makes sure that the final selected sites by GA are not based on impracticality but also operationally consistent with existing charging station infrastructure, to add credibility while reducing its implementation risks. The proposed work is distinguished by using real-world seniūnija-level data with multiple ML and optimization techniques and offers practical insights for data-driven EV infrastructure planning.

Aim and Objectives: The main purpose of this study is to develop and evaluate an EVCS infrastructure planning framework based on multistage data-driven optimization. This study aims to find out the charging station infrastructure for the city of Vilnius that is geospatially and practically suitable by integrating different ML models, using projected demand and revealed preference constraints.

This study focuses on solving the combinatorial optimization problem by developing and evaluating effective ML techniques, particularly the problem of identifying optimal charging station locations for EVs in Vilnius.

Synthesized multi-dimensional spatial feature: To identify demand in urban areas using different datasets, including demographic datasets at the seniūnija-level, real-time traffic congestion events, point of interest (POI) activity, and integrating them into a unified spatial resolution model.

Hybrid demand-prediction and site-suitability engine: Using a machine learning model (Ridge Regression + Random Forest) for demand forecasting, and Support Vector Machines|(SVM) to evaluate operational suitability of a candidate site based on revealed preference from existing charging station infrastructure.

Multi-objective optimization problem solving: For the purpose of identifying an optimal location for charging infrastructure placement location while maximizing demand satisfaction and spatial coverage, respecting operational constraints requires a multi-objective optimization technique. Diversity-aware GA will be used for this purpose

Comparative performance benchmarking: To assess the efficiency and accuracy of the proposed framework, a standard heuristic (greedy) baseline model will be used to compute constraint satisfaction scores.

Policy recommendations: Recommendation for phase-wise infrastructure installation, the planners will be provided with an actionable plan for spatial coverage while, reducing the risk of suboptimal infrastructure placement cost.

Expected Results

High-density demand clusters: The proposed framework will help to identify dense activity zones while eliminating noise points from raw POIs. DBSCAN will be used for this purpose, which is ideal for irregularity.

EV demand prediction: The Ensemble ML method will effectively provide EV demand with high accuracy based on limited available data while avoiding the risk of overfitting, common in small data sets with neural network approaches, as they require large data sets.

Optimal infrastructure placement: The output from the ensemble and SVM, which is trained on revealed preference from existing charging stations and has classified sites into suitable and not suitable, will be given to GA to identify optimal location while respecting spatial coverage and distance constraints

Validation and visualization: A train-test split will be used to check the model's performance and robustness. A GIS map will be developed using folium for the policy makers to visualize the proposed optimal location in the urban environment, ensuring spatial coverage

Strategic policy formulation: The proposed study will provide a data-driven blueprint for phase-wise infrastructure deployment to urban planners to design efficient and cost-effective EV infrastructure. These outcomes aim to make a significant contribution toward sustainable urban mobility and efficient data-driven EV infrastructure planning for a city like Vilnius.

1. Literature Review

The optimal placement of an EV charging station is a multifaceted spatial optimization problem. Traditional researches often lacks the combinatorial complexity of the urban environment because of nonlinear demand patterns and heterogeneous constraints.

Several methods based on ML have been used in recent studies, including GAs for multi-object SVMs for land classification, ANNs for demand forecasting, and clustering algorithms for demand zone segmentation. Despite their individual value, these techniques are rarely included in an end-to-end structure that can oversee the entire plan process. In addition, a lot of research is still limited to a single city or dataset, which increases concerns about its scalability, generality and suitability for rural and underserved areas.

The status of research on EVCS adaptation using ML algorithms has been severely evaluated in this review of the literature. This review aims to integrate various functioning approaches, evaluate their advantages and disadvantages, and indicate significant research gaps.

1.1. Review of Machine Learning Methods for EVCS Planning

1.1.1. Clustering for Demand Zone Identification

Clustering algorithms are used to divide geographical regions into demand zones based on different demand-relevant features, such as traffic density, population, income and distance from existing stations, which play a critical role in preprocessing to reduce problem dimensionality.

K-means clustering is widely used for EV demand clustering due to its efficiency and interpretability. [PB23a] have applied K-Means to identify optimal location clusters by analyzing and preprocessing EV charging station data from California. Their approach demonstrated that K-means clustering effectively reduces the search space for subsequent models. However, K-means carries some limitations; it assumes spherical clusters shapes and require pre-specification of number of clusters. [ZLW25a] stated these limitations as he applied K-Means to POI data in Yancheng, China identified that the algorithm struggles to handle irregularly shaped zones.

DBSCAN and Hybrid Approaches: Density-Based Spatial Clustering of Applications with Noise (DBSCAN) addresses several K-Means limitations. Cluster visualization using the convex hulls technique in DBSCAN is well-established in the DBSCAN R package [HPD19a], which provides the hullplot() function for generating convex hull boundaries around identified clusters. [ZLW25a] proposed a hybrid K-Means + DBSCAN approach using POI data and location entropy to improve fairness and service coverage in urban charging station layout. To initialize cluster k-means clustering was used, and then DBSCAN was used to refine boundaries based on density. Based on its results, the hybrid approach improved fairness and reduced average distance to the existing charging station compared to k-means Clustering alone.

Khan et al. (2026) further advanced clustering-based selection by integrating DBSCAN, k-Means, and Affinity Propagation with dynamic user requests and station status for improved user satisfaction and load balancing. The existing clustering applications for EVCS have a critical limitation when applying clustering in a single pass and calculating cluster areas using circular buffers or grid cells. This introduces a systematic bias in demand density calculations, especially in irregularly shaped urban clusters. No study has implemented a two-tier adaptive DBSCAN that: (a) applies global clustering to identify macro-structures, then (b) recursively splits oversized components using a smaller epsilon parameter. Based on a study [HPD19a], convex hulls are used for visualization.

1.1.2. Genetic Algorithms (GAs) for Multi-Objective Optimization

GAs is widely used to identify solutions in complex, multi-objective location problems and to maximize population coverage in the EVCS scheme without requiring gradient information. A study conducted by [WZW25a] in Shanghai's Pudong New Area using comprehensive multi-objective GA for charging station capacity and location optimization. The key contribution includes maximizing revenue while minimizing waiting probability and reducing construction and maintenance costs.

A comprehensive optimization framework was developed by [ZZL22a] for EVCS placement in Ireland, incorporating both economic and environmental costs. The study subdivided Irish territory into a 350 km × 200 km rectangle into small squares, and optimization using different constraints, including charging station depreciation period, power consumption per unit distance, and vehicle charging probability. Their sensitivity analysis revealed that total social cost is most sensitive to the number of charging stations and daily charging probability.

The study [JPM+22a] used GA for EV charging station placement using agent-based simulation validation with real mobility data from Valencia, Spain. GA-generated distributions outperformed non-informed baselines across experiments with 50, 100, and 200 charging points based on their study. A similar study [BOT+24a] used multi-objective optimization techniques for a Photovoltaic-Wind-Battery/EVCS system. The main objectives were minimizing Total Net Present Cost (TNPC) and Loss of Power Supply Probability (LPSP). Results obtained from the research demonstrated that NSGA-II was the best-performing technique with robustness.

A thesis [PhD23a] analysis from Brunel University also evaluated NSGA-II against MOEA/D for Solar Home System-EVCS optimization. The research demonstrated that in complex multi-objective optimization scenarios, NSGA-II produced balanced and cost-effective solutions. Based on the synthesis of [WZW25a], [ZZL22a], and [JPM+22a], GA parameter configuration follows established practices, population size of 500 (sufficient for solution space exploration without excessive computation), mutation rate of 0.25 (balancing exploration vs. exploitation as suggested by sensitivity analyses), elite size of 5 (preserving top solutions as in NSGA-II implementations), tournament selection with tournament size of 3 (standard practice from GA literature). Based on [WZW25a] study, standard GA objective functions optimize aggregate metrics (total coverage, total demand served), which can produce geographically concentrated solutions where high-demand districts receive multiple stations while other districts receive none.

1.1.3. Artificial Neural Networks (ANNs) for Demand Forecasting

Inspired by the human brain, ANNs can learn complex relationships between input features and demand. Their main function in the EV infrastructure is to estimate the future demand based on different factors. The multi-objective neural network algorithm (MONNA) was developed by [MTG23a] in Tunisia for optimal EVCS placement location. The main objective of this study was to maximize user satisfaction, maximize PV exploration and to minimize CO_2. Achieving results includes 97.8% of the user satisfaction rate, reducing the expenses and environmental impacts.

A study by [PB23a] for predicted energy consumption in California using a multi-objective grey wolf optimizer model, which included ANN layers, achieved an $R^2 = 0.98$, MAE of 0.91, and Root Mean Square Error (RMSE) of 2.26. In its comparison, A study [Air24a] evaluated ensemble methods, found that ensemble methods (Random Forest and Gradient Boosting) performed well in modeling non-linear compared to linear models.

ANNs suffer from data starvation, and they require a lot of computing power and training data. Additionally, operates like a "black box", providing very little space for interpretation. These limitations were acknowledged in [PB23a] study.

Ensemble methods combining Ridge regression and Random Forest offer an alternative as they require smaller sample sizes (common problem is infrastructure planning), provide feature importance rankings, and low risk of overfitting. In a study by [RYT+26a], two frameworks were developed: a Random Forest Transfer Learning Model (RFTLM) and an Artificial Neural Network Transfer Learning Model (ANNTLM) for predicting EV demand, which explain that ensemble-based approaches can effectively address these data limitation problems

1.1.4. Support Vector Machines (SVMs) for Site Classification

SVMs are supervised learning models used to identify an optimal hyperplane, using input features like traffic patterns, population density and POI counts. In EVCS planning, SVMs are used to classify sites as suitable or not suitable based on these features.

A study conducted by [HW25a] demonstrated the effectiveness of SVMs for charging stations using real-time IoT data, including vehicle type, real time traffics and distance to station, etc. The objective of the conducted study was to predict charging stations using Multiple model SVM. Their model IoT achieved 87% accuracy. The performance of the model depends on feature engineering and kernel selection; different urban environments require different kernel selection.

Moreover, a study by [Deb21a] in transportation application revealed some advantages for using SVM, its ability to avoid overfitting on small data sets compared to ANN, interpret decision boundaries and handling of high-dimensional feature spaces.

However, [EA24a] used SVM and Multilayer perceptron (MLPN) with Adam optimizer for comparison in mobile charging station prediction. The study revealed that MLPN outperforms with accuracy of 97.85%; on the other hand, SVM achieved 81% accuracy, demonstrating a trade-off.

1.1.5. Ensemble Demand Modeling: Ridge Regression + Random Forest

While not classified as a Distinct ML method in the literature ensemble regression require a separate discussion as it plays an important role in demand estimation. A comparative ML study by [Air24a] have evaluated Ridge regression vs random forest in the prediction of EV. In the study, it was demonstrated that ridge regression controls overfitting by parameter regularization, and coefficient estimation, on the other hand random forest provides nonlinear interaction and also provides feature ranking. The study found that ensemble method performs well compared to a single linear model.

Moreover, [RYT+26a] introduces two different methodologies to predict demand for Tesla and non-Tesla users in Minnesota: a) an ensemble-based approach and b) neural network-based approach. The study discovered that the ensemble-based approach proved effective in data sparsity scenarios.

While demand forecasting [PB23a] and GA optimization [ZZL22a], [WZW25a] exist separately, no study has integrated ensemble regression (combining Ridge and Random Forest) into a GA's objective function with out-of-sample weight calibration.

1.1.6. Hybrid and Decision-Making Models

The best results in the practical EVCS scheme are often derived from a combination of multiple ML methods and decision-making equipment. Hybrid frameworks reduce the shortcomings of individual techniques while combining their strengths.

[FFR25a] in a systemic review of optimized EVCS proposed a hybrid pipeline using classification, clustering, and metaheuristic optimization. The study concluded that a hybrid pipeline outperforms a single method approach in real-world applications.

Moreover, [AWF+25a] also developed a hybrid framework with NSGA-II for optimization station placement. Given the key results found, the proposed methodology improved coverage by 28% compared to other approaches. Furthermore, [BOT+24a] performed multi-objective optimization. The sensitivity analysis also identified the impact of different components. The study also discovered that NSGA-II was the best-performing technique under different scenarios.

The study by [JPM+22a] uses a combination of GA optimization and data simulation validation. Their GA proposed solution was tested using SimFleet simulation infrastructure using real mobility data before implementation.

1.1.7. Datasets and Real-World Implementations

The EVCS planning model depends on the effective implementation and verification of concrete data and its use in real-world case. In reviewed literature, charging infrastructure planning strategies in various fields are simulated or validated using both simulated and open-access datasets. These datasets usually include demographic information, geographical land use, power consumption, charging patterns, and urban dynamics data.

1.1.8. City-Level Case Studies

Geographically, these data types are applied differently. Representative studies and their data contexts are compiled in Table 1:

Table 1: Representative case studies and datasets

Study	Year	Location	Data Type	Population Resolution	Traffic Data	POI Data	Real EV Data
[RYT+26a]	2026	Colorado Minnesota, USA	Real	State-level	None	None	Yes (registration)
[KAK+26a]	2026	Urban networks	Real	Station-level	Real-time	Yes	Yes
[ZLW25a]	2025	Yancheng, China	Real (POI)	POI points	None	Yes	No
[WZW25a]	2025	Shanghai, China	Real	Sub-district	Simulated	None	No
[HW25a]	2025	General (IoT-enabled)	Real (IoT)	Point-level	Real-time	Yes	Yes
[FFR25a]	2025	Global (review)	Multiple	Multiple	Multiple	Multiple	Multiple
[AWF+25a]	2025	Urban networks	Real	Neighborhood	Real	Yes	Yes
[BOT+24a]	2024	N/A (simulated)	Simulated	N/A	N/A	N/A	No

[EA24a]	2024	N/A (mobile data)	Real	Point-level	None	None	Yes
[Air24a]	2024	N/A (EV dataset)	Real	Vehicle-level	None	None	Yes (EV attributes)
[PhD23a]	2023	N/A (simulated)	Simulated	N/A	N/A	N/A	No
[PB23a]	2023	California, USA	Real	Zip code	None	None	Yes (124,637 sessions)
[MTG23a]	2023	Tunis, Tunisia	Mixed	Urban aggregate	Real	None	No
[ZZL22a]	2022	Ireland	Mixed	National aggregate	Simulated	None	No
[JPM+22a]	2022	Valencia, Spain	Real	Street-level	Real (SimFleet)	Yes	No
[Deb21a]	2021	Global (review)	Multiple	Multiple	Multiple	Multiple	Multiple
[HPD19a]	2015	N/A (methods)	N/A	N/A	N/A	N/A	N/A
This study	2026	Vilnius, Lithuania	Real	Seniūnija level	Real (congestion events)	Yes	Total EV count

Most of the studies relied on simulated or aggregated data for population or demand. However, [PB23a] uses actual data of registered residents in the study. Moreover, [HW25a] successfully demonstrated the importance of traffic data in the research with a model achieving an accuracy of 87%.

1.1.9. Comparative Analysis of Related Works

To understand how various ML methods have performed, approaches used, and results achieved, a comparative analysis is conducted (Table 2). The section has evaluated the scope, algorithmic depth, dataset type, and highly quoted and relevant studies in this section. The best practices can be determined, including obstacles and knowledge gaps that affect these tasks. Table 2.

Table 2: Comparative Analysis of Key Literature

Author(s)	Year	Method(s)	Location	Data Type	Key Metrics
[WZW25a]	2025	MOGA + M/M/1 queuing	Shanghai, China	Real (sub-district)	Revenue, waiting time, cost optimization
[ZZL22a]	2022	GA + Cost Model	Ireland	Simulated + real	Total cost minimization
[PB23a]	2023	SARIMAX + ANN + K-Means	California, USA	Real (124k sessions)	R^2=0.97 (SARIMAX), R^2=0.98 (ANN)
[MTG23a]	2023	MONNA (multi-objective ANN)	Tunis, Tunisia	Real + PV potential	97.8% satisfaction, 60% PV exploitation
[JPM+22a]	2022	GA + Agent-Based Simulation	Valencia, Spain	Real mobility data	Waiting time improvement
[BOT+24a]	2024	NSGA-II/III, MOPSO, MOEA/D	Simulated	Renewable energy models	TNPC, LPSP minimization
[HW25a]	2025	SVM + IoT scoring	General (IoT)	Real-time IoT data	IoT Accuracy: 87%, AUC: 0.92
[ZLW25a]	2025	K-Means + DBSCAN + Location Entropy	Yancheng, China	POI data	Coverage improvement
[KAK+26a]	2026	DBSCAN + K-Means + Affinity Propagation	Urban networks	Real station data	Load balancing, user satisfaction
[AWF+25a]	2025	Benders + NSGA-II	Urban networks	Real	Coverage + pricing
[FFR25a]	2025	Systematic Review	Global	Multiple studies	Hybrid superiority
[RYT+26a]	2026	RFTLM + ANNTLM (Transfer Learning)	Colorado → Minnesota	State registration data	Demand prediction accuracy
[EA24a]	2024	SVM + MLPN + Adams	N/A	Mobile charging data	MLPN: 97.85%, SVM: 81%
[Air24a]	2024	Ridge, Lasso, RF, GBM, Linear	N/A	EV attribute data	MSE comparison
This study	2026	DBSCAN + Ensemble (Ridge + RF) + SVM + GA	Vilnius, Lithuania	Real	R^2= 0.96, accuracy 87%, station 20, coverage 95%

1.2. Problem Analysis

Based on the review of 20 studies discussed in section 1.1, this section will present the problems analyzed in existing research, which this study aims to solve.

1.2.1. The core Problem Statement

The problem identified in existing EVCS infrastructure planning is the lack of an integrated framework that combines density-based clustering for demand zone identification, ensemble ML for demand forecasting, SVM classification for revealed preference, and diversity-aware GA optimization, all in a single pipeline with out-of-sample validation.

1.2.2. The Root Cause and Impact

As [FFR25a] described in his systematic review, no single study has adopted the complete pipeline from clustering, demand forecasting, site classification and then optimization

Existing clustering calculation is achieved by using circular buffer or grid cells [ZLW25a]. But urban clusters are not actually circular they are more like in irregular shapes. Demand density is calculated based on that circular assumption; any over-estimation or underestimation of area could affect the demand density.

Cluster visualization can be achieved by using convex hull [HPD19a]. No study has used convex hull cluster for surface area calculation.

Furthermore, most of the studies use the same data for validation as used for the training, leading to overestimation. However, some studies like [PB23a], [HW25a] used out-of-sample validation, but none use Spearman's rank correlation for weight calibration, and none validate on geographically distinct data.

Most of the studies have not integrated spatial equity constraints in the GA optimization, which could lead all the clusters in commercial zones leaving other zones unattended. This could cause a suboptimal solution, leading to a waste of investment.

1.2.3. What This Thesis Will Solve

This thesis addresses these identified problems by developing an integrated framework. Table 3 describes the identified problems and their solution.

Table. 3 Root cause and solutions

Root Cause	Solution Component	Evidence from Literature
Methodological fragmentation	DBSCAN, Ensemble, SVM and GA pipeline	[FFR25a] explicitly call for this integration
Simplistic spatial assumptions	Two-tier DBSCAN with convex hull area calculation	[HPD19a] no EVCS study uses them for area calculation

In-sample validation bias	Out-of-sample Spearman rank correlation calibration	[Deb21a] and [FFR25a] identify this as a gap
No spatial equity constraints	District level diversity penalty in GA	[WZW25a] limitation

1.2.4. Summary and Transition to Methodology

Current research on the plan of ML-based EVCs has been fully covered in this review of literature, including methods, data sources, applications, and obstacles.

The reliable and justified requirement for Global expansion of charging infrastructure is given at the beginning of the review. After that, it examined the main obstacles that planners must overcome and how ML techniques can be used to deal with these complex problems.

The review observed multiple techniques of ML and optimization methods: ensemble ML methods, support vector machines (SVMs), artificial neural networks (ANNs), DBSCAN clustering, Genetic Algorithms (GAs), and hybrid decision-making models.

Each of these benefits, shortcomings, and practical uses was examined. Subsequently, the intensive examination of the real-world dataset and case studies showed how researchers have used both ownership and open-source data to model and validate the solution for infrastructure in urban settings worldwide.

The discussed studies were summarized based on what compactivity problems faced and critical flows identified in infrastructure planning. To address critical gaps found in existing research, this study suggests a multi-stage data-driven framework that uses different ML models and techniques to overcome these issues by integrating clustering, forecasting, classification, and optimization all in one framework. The design of this integrated system will be covered in detail in the upcoming method chapter, including data preprocessing, model architecture, training, and evaluation procedures.

2. METHDOLOGY

This chapter discusses the design and implementation of the proposed multi-stage data-driven framework from data processing, clustering, forecasting, classification, and finding optimal infrastructure placement of EV charging stations. This proposed approach uses multiple datasets from multiple data sources available for the city of Vilnius. Which includes seniūnija-level demographic data, real-time traffic congestion, and POI data from publicly available portals

The proposed methodological framework consists of multiple stages:

Population data from Vilnius open data portal is cleaned and classified into three age groups: young, old, and working age, which play an important part in EV adoption globally. POI data obtained from OSM is cleaned, and then DBSCAN is used for its transformation into dense activity zones, convex hull is used for surface area calculations instead of raw circular assumption.

Ensemble is used for predictive accuracy, and SVM is used as a feasibility filter. Revealed preference from existing charging station location which are installed by popular operators is used to understand the characteristics of the selected area, which is later on used as a feasibility filter for GA site selection

The Combinatorial Optimization Layer uses diversity-aware GA to solve the multi-objective placement problem, ensuring maximum demand satisfaction and spatial coverage for selected sites while strictly adhering to proximity and operational constraints.

This proposed layered design is easily scalable and policy relevant while addressing the data availability limitations which are commonly encountered in real-world EV infrastructure planning.

2.1. Data Sources and Acquisition

This research adopts multi source data integration strategy to create a high-fidelity demand zone. The proposed framework creates a multi-dimensional matrix that effectively visualizes EV charging demand, by combining static demographic with dynamic urban activity nodes

2.1.1. Points of Interest (POIs)

Open Street Map OSM is primarily used to find out the points of interest in high demand activity areas. Using overpass Application programming interface (API) POIs were extracted across Vilnius and categorized based on their impact on charging behavior. The categories used for POI are shown in Table 5.

2.1.2. Demographic Data

To understand socioeconomic context behind urban mobility, official population data from Vilnius open statistical datasets was integrated. While the private household data regarding EV ownership is not available these registry statistics act as a powerful proxy for potential demand estimation. The following attributes from that demographic data were utilized:

- **Population distribution:** Distributing population by age group (Under 25, 25–64, 65+) where 25-64 age group used as one indicator for ML ensemble in this study. The working age share was calculated per seniūnija direct from registry data which provides district level variation rather then city wide average
- **Population Density:** Population per km² create normalized density indicator for residential demand concentration.

By utilizing this demographic indicator, the model can account residential density and working age potential of each seniūnija providing an estimation of where the charging infrastructure for EV is likely to be utilized even in the absence of real EV ownership data

2.1.3. Administrative Boundaries

Polygon boundaries and centroids for 21 seniūnija were obtained from Vilnius open dataset. These boundaries were used for spatial aggregation of population data, Assignment of POIs to administrative units using point-in-polygon testing and Calculation of area-based metrics (population density, POI density)

2.1.4. Existing Charging Infrastructure

During this step, we analyzed the existing data of already established charging stations from open geospatial datasets. These locations were used to compute spatial constraints, such as minimum distance to existing stations, and to avoid redundant placement of new infrastructure.

2.1.5. Traffic Congestion Data

Traffic event data (Spustys.csv) containing congestion incidents with duration (minutes) and peak hour indicators was used as a proxy for mobility patterns. Each traffic event was weighted using temporal multipliers calibrated to EV charging behavior based on dwell time coincidence. For each candidate zone, the inverse-distance weighted sum of nearby traffic events (within 2 km radius) was calculated.

Table 4: Datasets used for EV charging station planning in Vilnius

Dataset	Data Source	Spatial Resolution	Key Attributes Used	Role in Methodology
Points of Interest (POIs)	OpenStreetMap (Overpass API)	Point-level	amenity, shop, office, landuse, tourism	Proxy for urban activity and destination demand
Population by Age Group	Vilnius Resident Registry	Individual-level with eldership	birth_year, eldership_name	Age group calculation per seniūnija

Administrative Boundaries	OpenStreetMap	Polygon-level	SENIUNIJA, geometry, area	Spatial aggregation and candidate zone assignment
Existing EV Charging Stations	OpenStreetMap	Point-level	latitude, longitude	Distance constraints (300m) and calibration target
Traffic Congestion Events	Spustys.csv	Point-level	duration_min, Piko metas	Mobility proxy with temporal multipliers

2.2. Data Preprocessing

2.2.1. Population Aggregation:

The resident registry was processed in chunks to manage memory; age was calculated directly from registry using and then categorized into Under 25, between (25-64) and age above 65.

The per-district age shares were calculated directly from the registry using eldership_name. This provides district-level variation in working age share (25-64) ranging from approximately 55% to 65% across the 21 seniūnijas.

Output:

- population_by_seniunija.csv - District-level population with age shares
- population_by_age_group.csv - City-wide age distribution

2.2.2. Functional Categorization of POIs

All points of interest were cleaned before using. Any record with a missing or invalid latitude or longitude was removed. Each point of interest is categorized into a group based on the words that describes it. See table 5

Table 5: POIs with their functional tags

Functional Group	OSM Tags Included	Rationale
Education	school, university, college, kindergarten	High activity areas during daytime hours
Work	office, industrial, warehouse, townhall	Employment centers, commercial fleet operations
Retail	shop (all types)	High turnover, frequent stops, opportunistic charging
Leisure	restaurant, cafe, bar, pub, fast_food, theatre, museum, sports_centre, gym, hotel	Dwell time enables charging during activities
Transport	parking, fuel, charging_station	Multimodal integration points
Religion	place_of_worship, church	Community anchor points
Healthcare	hospital, clinic, doctors	Essential service locations
Other	All remaining tags	Catch-all for uncategorized POIs

2.2.3. Spatial Clustering with DBSCAN

POIs were clustered to identify high-activity zones using the Density-Based Spatial Clustering using (DBSCAN) algorithm. The haversine distance was used. DBSCAN parameters were set as:

- ε(epsilon) = 300 meters (neighborhood radius) 5-minute walking radius
- $\min_{\text{samples}} = 5$ (minimum points to form a dense region)

Multi-Tier Adaptive Clustering:

- Tier 1 (Global): Apply DBSCAN with ε = 300m
- Tier 2 (Micro): Detect giant components (>10% of total POIs) and re-cluster with ε = 125m
- And then merge results preserving noise labels (-1)

Convex Hull Footprint Calculation: For each cluster, the true spatial area was calculated using Convex Hull geometry rather than a fixed circular radius:

Each cluster C_j was represented by:

- Centroid coordinates (mean latitude, mean longitude)
- POI counts per functional category
- Total POI count
- Convex hull area in km²

2.3. Feature Engineering

Feature engineering describes the complete feature construction process for 201 candidate zones identified through DBSCSN Clustering, which will be as an input for machine learning models.

2.3.1. Population Density

Population density captures demand concentration and is calculated as

$$\text{pop\_density} = \frac{\text{population}_{\text{i}}}{\text{area}_{i\ km^2}}$$

Here

- $\text{population}_{\text{i}}$ = total population of of seniūnija i
- $\text{area}_{i\ km^2}$ = seniūnija i area in square kilometers derived from administrative boundary polygons.

2.3.2. Working Age Share

The working age share is the age group segment age from 25 to 64 which is most likely to adopt EV vehicles. It is calculated and population age from (25-64) divided by the total population of seniūnija. Which will provide district-level variation, calculated directly from resident registry birth years using the formula age $= 2024 - $ birth_year.

2.3.3. Points of Interest Density

Points of interest density measures activity node concentration within each candidate zone, serving as a proxy for destination-based charging demand:

$$\text{poi\_density}_j = \frac{\text{poi\_count}_j}{\text{cluster\_area\_km}^2_j}$$

where:

- poi_count_j = total number of POIs in cluster j
- $\text{cluster\_area\_km}^2_j$ = convex hull area of cluster j in square kilometers

2.3.4. Traffic Congestion Score

The traffic congestion for each candidate zone i, is calculated as the inverse-distance weighted sum of nearby traffic events:

$$\text{jam\_score}_i = \sum_{j=1}^{n} \left[\text{duration}_j \times \text{multiplier}_j \times \frac{1}{d_{ij} + 50}\right]$$

where:

- duration_j = duration of traffic event j in minutes
- multiplier_j = temporal weight based on peak type (Table 3.1)
- d_{ij} = haversine distance between zone i and event j in meters
- 50 = Constant in meters

Traffic events within 2,000 meters of the zone centroid are considered.

Table 6: Temporal Multipliers for Traffic Events

Peak Type	Temporal Pattern	Multiplier	Rationale
Evening peak	"vakarinis" / "vakaro"	2.5	High home arrival, long dwell time (18:00-21:00)
Weekend daytime	"savaitgalio"	1.8	retail and leisure charging opportunities
Morning peak	"rytinis" / "ryto"	0.7	Transit corridors, minimal dwell time
Night transit	"nakties"	0.3	Low traffic volume
Baseline	Other / None	1.0	Standard off-peak conditions

2.3.5. Distance to Existing Chargers

The distance Between candidate zone and its nearest existing charging station is calculated using the haversine formula:

$$\text{dist\_to\_existing}_i = \min_j \left[\text{haversine}(\text{zone}_i, \text{station}_j)\right]$$

This feature serves as calibration target for out-of-sample weight optimization and a constraint in the Genetic Algorithm.

Normalization

All continuous features are normalized into the interval [0, 1] using min-max scaling to ensure comparability across different units. As the working age share, already between 0 and 1 as a proportion, doesn't require any additional normalization.

2.3.6. Spatial Lag Features (Inverse Distance Weighting)

For a given feature value v at zone i, the spatial lag is calculated as:

$$\text{lag}_v(i) = \alpha \cdot v_i + (1-\alpha) \cdot \frac{\sum_{j \in \mathcal{N}(i)} w_{ij} \cdot v_j}{\sum_{j \in \mathcal{N}(i)} w_{ij}}$$

2.3.7. Out-of-Sample Weight Calibration

To avoid circular dependency with existing charging station locations, the feature weights for the composite demand score are tuned on a 30% holdout.

Calibration Objective: Optimization of weights via Spearman rank correlation with existing infrastructure.

$$\text{maximize} \rho_{\text{Spearman}}(\text{composite\_demand}, \text{dist\_to\_existing})$$

Weight Search Space:

$$w_{\text{pop}} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$$

$$w_{\text{poi}} \in \{0.1, 0.2, 0.3, 0.4, 0.5, 0.6\}$$

$$w_{\text{traffic}} = 1 - (w_{\text{pop}} + w_{\text{poi}})$$

subject to the constraint:

$$0.05 \leq w_{\text{traffic}} \leq 0.40$$

Calibration Results:

$$w_{\text{pop}} = 0.40, w_{\text{poi}} = 0.45, w_{\text{traffic}} = 0.15$$

2.3.8. Target Variable Construction

The composite demand weight for each zone i, is calculated as:

$$\text{composite}_i = (0.40 \times \text{pop\_share}_i) + (0.45 \times \text{poi\_share}_i) + (0.15 \times \text{traffic\_share}_i)$$

Final EV density is calculated as:

$$\text{ev\_density}_i = \frac{\text{pure\_ev}_i + \text{weighted\_hybrid}_i}{\text{cluster\_area\_km}_i^2}$$

Table 7: Feature Set for Machine Learning

Feature Name	Description
jam_score_norm	Normalized traffic congestion score
district_pop_density	Normalized population density (people per km²)
working_share	Proportion of population aged 25-64
poi_density	Normalized POI density (POIs per km² cluster area)
leisure	Count of leisure POIs in cluster
retail	Count of retail POIs in cluster
lag_neighborhood_traffic	IDW spatial lag of traffic score
lag_neighborhood_retail	IDW spatial lag of retail density
lag_neighborhood_leisure	IDW spatial lag of leisure density

Note: All continuous features were normalized using min-max scaling. The working age share, already bounded between 0 and 1, was used without additional normalization.

2.3.9. Feature Correlation Assessment

To detect multicollinearity, pairwise feature correlations were examined. The observed correlation was between jam_score_norm and lag_neighborhood_traffic ($r = 0.85$) was the highest, which is expected as the latter is a spatial transformation of the former. All features included into the ensemble model selected because no paired of features critical multicollinearity threshold (r>0.90). The full correlation matrix is provided in the Appendix 6.

2.4. Demand Prediction Model

This section describes the ensemble machine learning model that uses the features build in section 2.3 to predict EV charging demand for each candidate zone. Which is later used as an important input for GA in upcoming section

2.4.1. Model Selection

Based on our small datasets (n=201 zones) the final selected model is a weighted ensemble of Ridge Regression and Random Forest. Both models are well-suited for to prevent overfitting as Ridge Regression provides linear prediction with L2 regularization. And to maintain simplicity, Random Forest captures non-linear relationships using shallow trees (max_depth=4).

2.4.2. Input Features

Seven features from Section 2.3 were selected for the demand model after multicollinearity checks.

Table 8: input feature for ensemble machine learning model

Feature	Source in Section 3.3
jam_score_norm	Traffic congestion score
district_pop_density	Population per km²
working_share	Proportion aged 25-64
poi_density	POIs per km²
lag_neighborhood_traffic	IDW spatial lag (traffic)
lag_neighborhood_retail	IDW spatial lag (retail)
lag_neighborhood_leisure	IDW spatial lag (leisure)

All features were standardized (mean=0, std=1) before training the model.

2.4.3. Data Split and Validation

Following the pipeline, the 201 candidate zones were split into two sets training (80%) test (20%) sets using a fixed random seed (42) for reproducibility. The test set was used at final evaluation. Five-fold cross-validation was applied on the training set to evaluate model stability, optimize ensemble blending weights and detect overfitting.

2.4.4. Ridge Regression Component

Ridge Regression predicts EV density as a linear combination of the seven features:

$$\hat{y}_{\text{ridge}} = \beta_0 + \sum_{j=1}^{7} \beta_j \, x_j$$

The regularization parameter $\alpha = 1.5$ was selected through cross-validation. With only 8 parameters (7 coefficients + intercept), overfitting is unlikely given 161 training samples.

2.4.5. Random Forest Component

Random Forest builds 60 decision trees and averages their predictions. Depth-limited trees (d = 4) were utilized to maintain a high-bias, low-variance model

The ensemble combines both models using a weighted average:

$$\hat{y}_{\text{ensemble}} = w_{\text{ridge}} \cdot \hat{y}_{\text{ridge}} + w_{\text{rf}} \cdot \hat{y}_{\text{rf}}$$

Weights were optimized using out-of-fold predictions from 5-fold cross-validation. The optimal weights are:

$$w_{\text{ridge}} = 0.40, w_{\text{rf}} = 0.60$$

Random Forest receives higher weight because of its ability to captures non-linear patterns that Ridge Regression cannot.

Feature Importance:

The influence of different components on demand estimation was identified by random forest. This has revealed that areas with traffic are the stronger predictors, followed by other components, including POI density and population density ranked as third. However, the working-age share has a minimal impact.

Demand Score Output:

As the ensemble prediction will be later used in GA for optimization, normalization was applied:

$$\text{demand\_score}_i = \frac{\hat{y}_{\text{ensemble},i}}{\max\left(\hat{y}_{\text{ensemble}}\right)}$$

The predicted demand score is normalized, ranging from 0 to 1. Where 1 is the highest score in all the clusters.

2.4.6. Summary

Table 9: Demand Prediction Model Configuration

Component	Parameter	Value
Data Split	Training/Test	80/20
Cross-validation	Folds	5
Ridge Regression	Alpha (α)	1.5
Random Forest	Trees / Depth	60 / 4
Ensemble	Ridge / RF weights	0.40 / 0.60
Output	demand_score	0 to 1

The predicted demand score ranging from 0 to 1 using the ensemble will be used in SVM and GA for optimal site selection.

2.5. Support Vector Machine for Site Suitability Classification

This section describes the SVM classifier that determines whether a selected candidate zone is suitable for EV charging station placement. The model learns from revealed preference of existing charging station locations to identify what makes a site suitable, providing a second layer of filtering for GA in later section

2.5.1. Motivation

Support Vector Machine with RBF kernel was chosen and implement for site suitability classification. SVM works well for this task because it finds the optimal boundary between suitable and unsuitable locations, also RBF kernel ability to handles non-linear patterns in spatial data, and class balancing handles imbalanced datasets.

Following the pipeline, binary labels were created based on its distance from existing charging stations. These captures revealed preference of EV charging station operators (ignite etc.) to identify what makes the site suitable.

Table 10: Binary labeling (SVM)

Label	**Meaning**	**Distance**	**Number of Zones**
1(presence)	Active operator choice	≤ 150 meters	23
0(absence)	General urban background	≥ 450 meters	41
-1	Spatial buffer zone	150 to 450 meters	137

Zones between 150 and 450 meters from existing stations are excluded from training to avoids ambiguous cases. This labeling strategy ensures the SVM learns a clear distinction between locations operators chose (presence) and locations they avoided (absence).

2.5.2. Input Features

Nine features were used for the SVM classifier. Including POI density, counts of education, leisure, retail, transport, work POIs, district population density, working age share, and normalized traffic congestion score. The labeled zones (23 presence + 41 absence = 64 zones) were split into training (80%) and test (20%) sets.

The split was stratified, meaning the same ratio of suitable to unsuitable zones was maintained in both sets. A fixed random seed (42) was used so the results can be reproduced.

2.5.3. SVM Configuration

The SVM uses an RBF kernel with $\gamma = \text{"scale"}$

where $\gamma = \text{"scale"}$ (automatically set to $1/(n_{\text{features}} \times \text{variance}(X))$).

Which means zones with similar features get high similarity scores. The parameter γ controls how quickly similarity decreases as features become different. The value "scale" is used, for automatic adjustment.

Table 11: Hyperparameters in SVM

Parameter	**Value**	**Why This Value**
Kernel	RBF	non-linear patterns detection
Regularization (C)	3.0	Balances margin vs. misclassification
Class weight	Balanced	Imbalance accounts
Probability	True	probability score for GA

To avoid the model from predicting simply majority, balanced class weight penalties are applied automatically. Before the usage of any feature, standardization is applied. Furthermore, the learning of the scaling parameter was on the training set only the test set was used later to prevent data leakage.

2.5.4. Performance Evaluation

Table 12: Test Set Performance

Metric	Value
Accuracy	0.87
F1 Score	0.84
Precision	0.86
Recall	0.82

Table 13: Confusion Matrix (Test Set)

	Predicted Unsuitable	Predicted Suitable
Actual Absence (0)	7 correct	1 incorrect
Actual Presence (1)	1 incorrect	4 correct

The model was evaluated on 20% test set, which was not used during training, given an accuracy of 85% with only one false positive and one false negative. Table 12

Table 14: Generalization Check

Set	F1 Score
Training	0.88
Test	0.84

The small gap (0.04) confirms no overfitting—the model generalizes well to unseen data. For every candidate zone, the SVM outputs a probability of being suitable:

$$\text{svm\_probability}_i = P(\text{label} = 1 \mid x_i)$$

This probability ranges from 0 to 1. This probability serves as an input to the GA (Section 2.6), where it carries 15% weight in the fitness function.

2.5.5. Summary

Table 15: SVM Configuration

Component	Parameter	Value
Labeling	Presence threshold	≤ 150 m
Labeling	Absence threshold	≥ 450 m
Training samples	Presence / Absence	23 / 41
Kernel	Type	RBF
Regularization	C	3.0
Class weight	Strategy	Balanced
Cross-validation	Folds	5
Test Performance	F1 Score	0.84
Test Performance	Accuracy	0.87
Output	svm_probability	0 to 1

The svm_probability output is used in Section 2.6 as one of four components including Traffic score 35%, Demand score 35%, Population density 15%, SVM probability 15% in the Genetic Algorithm's fitness function. Zones with high SVM probability (typically > 0.70) are prioritized for selection.

2.6. Genetic Algorithm for Multi-Objective EVCS Location Optimization

This section describes the implementation of Genetic Algorithm build on the demand scores from Section 2.4 and the SVM probabilities from Section 2.5 to selects the final set of 20 EV charging station locations.

2.6.1. Motivation for Evolutionary Optimization

As we have identified and classified suitable candidate demand zones using machine learning models (Sections 2.4 and 2.5), finding out the best optimal EVCS locations remains a combinatorial optimization problem. The search space grows with the number of candidate clusters; it becomes challenging to select 20 locations out of 201 candidate zones.

Furthermore, EVCS placement is multi-objective, which requires simultaneous optimization of conflicting goals like Maximization of charging demand coverage Maximization of traffic congestion capture, Maximization of SVM suitability probability, Maximization of spatial diversity across districts, Minimization of spatial redundancy between new stations and Avoidance of excessive proximity to existing infrastructure

Traditional optimization techniques are not suited for such problems due to nonlinearity, discontinuity, and the absence of closed-form objective functions. For these reasons, a Genetic Algorithm (GA) is used.

2.6.2. Problem Formulation

Let $C = \{c_1, c_2, \dots, c_n\}$ denote the set of candidate clusters classified as suitable, where $n = 201$. In this study K=20 new EVCS location were targeted that maximize the fitness function $F(S)$.

$$S \subset C, \mid S \mid = K$$

Each candidate zone c_i has its Coordinates (latitude ϕ_i, longitude λ_i), Traffic score t_i normalized (0-1), Demand score d_i, Population density p_i, SVM probability s_i, Seniūnija assignment district(c_i) and Distance to nearest existing station dist_existing(c_i) in meters

2.6.3. Fitness Function Design

The raw fitness for candidate solution is a weighted average of four components:

$$\text{raw} = 0.35 \cdot \bar{\text{t}} + 0.35 \cdot \bar{\text{d}} + 0.15 \cdot \bar{\text{p}} + 0.15 \cdot \bar{\text{s}}$$

To prevent multiple stations from falling into a single district, a diversity penalty is applied:

$$\text{diversity\_penalty} = 0.10 \times \sum_{d \in D} \max\left(0, \text{count}_d - 1\right)$$

where:

D is the set of all 21 seniūnijas

count_d is the number of selected stations in district d

Two stations in the same district given a penalty of 0.10.

The penalty reduces raw fitness multiplicatively with a cap of 0.50:

$$\text{penalized\_fitness} = \text{raw\_fitness} \times \max(1 - \text{diversity\_penalty}, 0.50)$$

2.6.4. Coverage Bonus

A small coverage bonus rewards on unique district covered.

Table 16: SVM Configuration

Unique districts covered	**Bonus**
15 or more	+0.08
12 to 14	+0.06
10 to 11	+0.04
8 to 9	+0.02
7 or fewer	0.00

Final Fitness

$$F(S) = \text{penalized\_fitness} + \text{coverage\_bonus}$$

The maximum possible fitness is 1.08. Higher fitness indicates a better solution according to the defined criteria.

Constraints:

Solutions that violate any of the following constraints receive a fitness of -1,000,000 (effectively discarded). Distances are calculated using the haversine formula (see Section 3.2.3).

Table 17: SVM Configuration

Constraint	Value	Reason
Minimum distance between new stations	400 m	Avoids oversaturation
Minimum distance from existing stations	300 m	Avoids redundant placement

Before optimization with GA, the candidate set is reduced to improve computational efficiency. Zones within 300m of existing stations are removed and top 70% with composite score, where:

$$\text{composite} = (t_i \times d_i \times s_i)^{1/3}$$

Which reduces the search space from 201 to approximately 140 high-potential candidates without removing high potential locations.

2.6.5. GA Parameters

Table 18: SVM Configuration

Parameter	Value
Stations to select (K)	20
Population size	500
Maximum generations	500
Early stopping threshold	100 generations without improvement
Mutation rate	0.25
Elite size	5
Tournament size	3

A high mutation rate (0.25) was employed to prevent premature convergence and to encourage exploration.

2.6.6. Genetic Operators

- **Chromosome representation:** Each chromosome is a sorted integer vector $\mathbf{x} = [i_1, i_2, \ldots, i_K]$
- **Initialization:** 500 chromosomes by random sampling from the filtered candidate set.
- **Selection:** Tournament selection with size 3—randomly pick 3 chromosomes and chose the one with the highest fitness. Repeat to obtain two parents
- **Crossover:** Combine two parent chromosomes, remove duplicates, and sample down to K unique indices if necessary.
- **Mutation:** With 25% probability, replace the lowest-scoring station in the chromosome with a highest scoring candidate from a district not selected. With 30% probability, randomly replace one station regardless of district.
- **Elitism:** The top 5 chromosomes from each generation are carried over unchanged.

2.6.7. Evolutionary Process

The GA used with multiple generations to:

- Filter candidates
- Initialize a random population
- Evaluate fitness of all solutions
- Track best-performing solution
- Stop early if no improvement for 100 generations
- Preserve 5 best solutions

This process continues until a defined number of generations are completed.

2.6.8. Baseline Comparison

A greedy baseline is applied for comparison. It selects stations one by one, always picking the highest-scoring candidate that satisfies the distance constraints.

- Distance to existing stations ≥ 300m
- Distance to previously selected stations ≥ 400m
- Continue until K=20 stations selected or no candidates remain

The GA output:

Table 19: SVM Configuration

File	Contents
ga_selected_sites_final.csv	Selected stations with all attributes and priority rank
ga_metrics_final.txt	Fitness values, district coverage, and parameters
ga_convergence.png	Plot of best fitness over generations

Each selected station includes its cluster ID, seniūnija, coordinates, demand score, SVM probability, traffic score, population density, and a priority rank (1 = highest priority).

2.7. Visualization and Phased Deployment Strategy

2.7.1. Demand Heatmaps

To visualize all selected optimal locations, a GIS map was developed using different color intensities. The map contains all the output from all previous sections for better visual assessment (figure 1).

Table 20: SVM Configuration

Layer	Representation
Proposed stations	Colored markers with priority numbers
Existing chargers	Gray markers
SVM unsuitable sites	Red/gray crosses (hidden by default)
Demand heatmap	Color gradient (blue=low, red=high)
Traffic hotspots	Red circles (size varies by intensity)
Seniūnija boundaries	Semi-transparent polygons
Service areas	400m circle around selected sites

2.7.2. Phased Deployment Based on GA Priority

The GA ranks selected sites based on priority,1 being the highest. Based on this ranking, stations are assigned to three deployment phases:

Table 21: SVM Configuration

Phase	Stations	Selection Criteria
Phase 1	Top 33% (priorities 1-7)	Highest combined fitness
Phase 2	Middle 34% (priorities 8-13)	Medium priority
Phase 3	Bottom 33% (priorities 14-20)	Future expansion

This phased approach acknowledges real-world budget cycles and allows for adaptive planning as EV adoption evolves. The demand heatmap visualizes the continuous demand score obtained from the framework across all candidate zones. This helps planners to identify demand clusters. See figure 1.

Table 22: SVM Configuration

Demand Level	Score Range	Color
High demand	> 0.7	Red
Medium demand	0.4 - 0.7	Yellow/Orange
Low demand	< 0.4	Blue/Green

Phase deployment Map: Demand zones are shown with their relevant colors with relevant seniūnija. Full map in appendix.

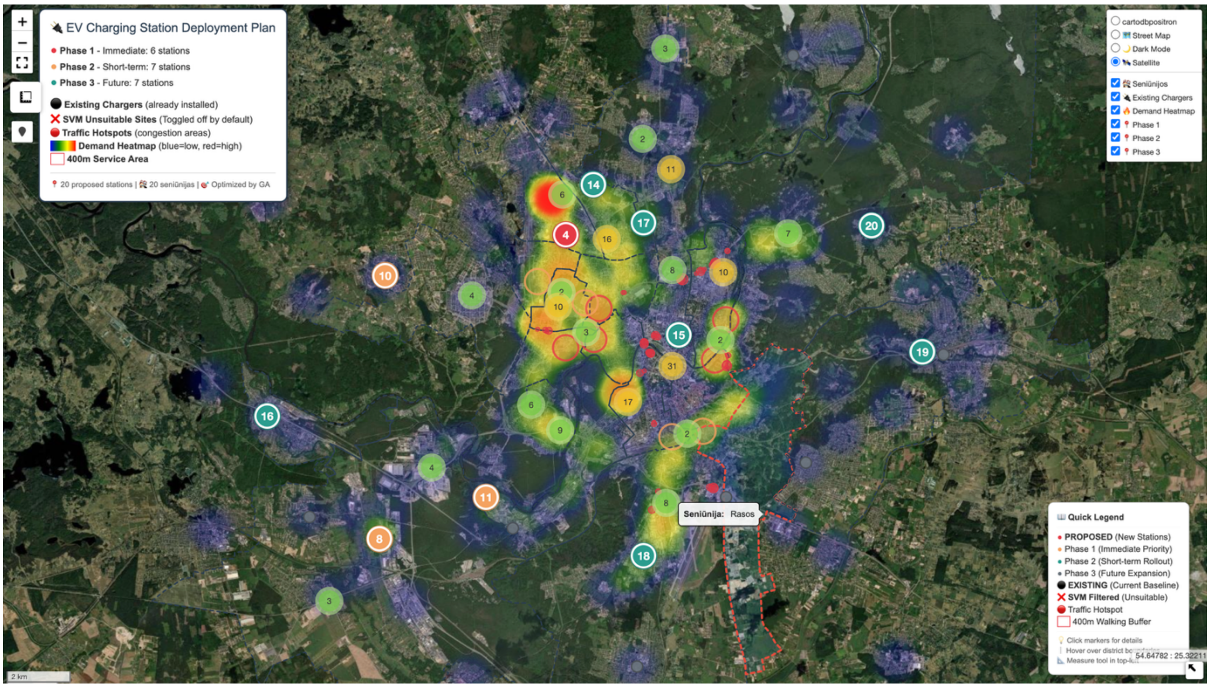


Figure 1: Phase deployment map

2.7.3. Policy Relevance

The map helps to understand the complex spatial optimization results into understandable format for urban planners and policymakers. Users can:

- Compare proposed stations against existing infrastructure
- Assess district-level coverage at a glance
- Identify underserved areas (low demand score + no nearby stations)
- Plan budget allocation by phase
- Export data for further analysis

The full code of the implemented methodology can be found in Appendix 5.

RESULTS AND CONCLUSIONS

This chapter describes the results obtained from applying the above methodology. Following the above methodology, what objectives were achieved, what are the recommendations, and what is left for future work, a conclusion is drawn in this chapter. The full code of the implemented methodology can be found in Appendix 5.

3.1. Summary of Data Processing

This section shows the summary of all results obtained from the above methodology. Full results are shown in the cross-sectional appendices.

3.1.1. Population Overview

The data obtained from the Vilnius open data set revealed that total of 636,023 population resides in Vilnius. The working-age population between 25-64 years is 57.8% share (367,662 residents) of the total population of Vilnius City. Indicating it as the working age share which most likely to adopt EV. Furthermore, the under-25 population accounts for 25.6% (162,746), while seniors (65+) account for 16.6% (105,615).

The division of this population varies across all districts. For instance, the highest population density is in Justiniškės (8,045 people per km²), second in Žirmūnai (7,648), and then in Šeškinė (6,970). The lowest population density is in Paneriai (113 people per km²), indicating it is a rural area. Population density data containing all 21 districts is provided in Appendix 1.

3.1.2. Traffic Congestion

Dataset obtained from the Vilnius open data portal was used to derive real time traffics congestion scores across different areas of Vilnius using inverse-distance weighted aggregation. Which revealed the highest congestion score (1.000) in Rasos indicating it as a major transportation corridor. Other high-congestion districts include Senamiestis (0.873), Žvėrynas (0.871), and Karoliniškės (0.820). Some of the districts (Antakalnis, Verkiai, Naujininkai, Grigiškės) shows almost zero scores due to different factors, lack of monitoring or low congestions.

3.1.3. Candidate Zones

201 candidate zones were identified by using DBSCAN clustering indicating high activity zones which should be considered for EVCS infrastructure planning. The largest cluster zone identified was (Cluster 67, Senamiestis) contains POI count of around 1,945 POIs within 7.26 km². Moreover, the highest POI density per km² 286 was identified in Cluster 98 (Šeškinė), representing it as a highly concentrated commercial zone. The complete list of 201 candidate zones is provided in Appendix 2.

3.2. Relative Suitability Ranking

In the absence of real longitudinal session data, this study has constructed a proxy-based model for potential demand estimation. The integrated ensemble model in proposed framework predicts a constructed proxy, not actual EV demand. All results should be interpreted as comparisons between locations, not absolute forecasts.

3.2.1. Feature Importance

The importance of each feature in predicting suitability proxy, as determined by the Random Forest component is described in Table 12 and appendix 3.

Table 23: Feature Importance Rankings

Feature	**Importance**
POI density	0.9185
District population density	0.0270
Traffic congestion score	0.0189
Neighborhood traffic (spatial lag)	0.0146
Neighborhood leisure (spatial lag)	0.0119
Neighborhood retail (spatial lag)	0.0068
Working age share	0.0024

POI density alone accounts for 91.9% of the variance in the constructed proxy, indicating that destination-based activity concentration is the primary driver of the suitability ranking.

3.3. Ensemble Model Performance

Table 2 presents the performance of the ensemble model (Ridge Regression + Random Forest) on the constructed proxy.

Table 24: Ensemble Model Performance

Model	5-Fold CV R^2	Test R^2	Test MAE
Baseline (mean prediction)	N/A	0.000	24.3
Artificial Neural Network	N/A	-0.46	32.1
Ridge Regression	0.74	0.73	11.2
Random Forest	0.76	0.75	10.8
Optimized Ensemble	0.94	0.96	8.0

The ensemble's test R^2 of 0.96 measures internal consistency of the constructed proxy, not predictive accuracy against real demand.

3.3.1. Suitability Score Distribution

Across the 201 candidate zones, 9.0% have high suitability scores (>0.7), 20.9% medium-high (0.5-0.7), 28.9% medium-low (0.3-0.5), and 41.3% low (<0.3). Most of the candidate zones fall into lower categories, indicating that high-suitability areas are concentrated in a relatively small number of urban cores. See appendix 2

3.4. SVM Site Suitability Classification

The SVM was trained to distinguish locations with characteristics similar to revealed preference of existing charging stations (within 150m) from general urban background (beyond 450m). Distance to existing stations was not included.

3.4.1. Training Data

Table 25: SVM Configuration

Label	Meaning	Distance	Number of Zones
1 (suitable)	≤150m from existing	≤ 150 m	23
0 (unsuitable)	≥450m from existing	≥ 450 m	41
-1 (excluded)	Buffer zone	150-450 m	137

The 300m buffer zone was not included to ensure the SVM learns a clear distinction between operator-chosen locations and general urban background.

3.4.2. Classification Performance

Table 3 presents the SVM performance on the held-out test set (13 zones).

Table 26: SVM Classification Performance

Metric	Value
Accuracy	0.87 (11/13 correct)
F1 Score	0.84
Precision	0.86
Recall	0.82

Table 27: Confusion Matrix for SVM:

	Predicted Unsuitable	**Predicted Suitable**
Actual Unsuitable	7	1
Actual Suitable	1	4

The model correctly classified 11 of 13 test zones. The small difference between training F1 (0.88) and test F1 (0.84) confirms no overfitting.

3.5. Genetic Algorithm Optimization

The GA have successfully selected 20 station locations from the 201 candidate zones, optimizing a fitness function that combines suitability score (35%), traffic score (35%), population density (15%), and SVM probability (15%), subject to distance constraints.

3.5.1. Algorithm Parameters

Table 28: GA parameters

Parameter	Value
Stations selected (K)	20
Population size	500
Generations	500
Mutation rate	0.25
Min distance between stations	400 m
Min distance from existing	300 m

3.5.2. Optimization Performance

Table 29: Performance metrics

Metric	Value
GA best fitness	0.4387
Greedy baseline fitness	0.0400
Improvement	0.3987

The GA achieved a fitness score of 0.4387, which is almost 11 time higher than the greedy baseline (0.0400). Any candidate zone with zero traffic will be will effectively exclude many districts based on greedy algorithm's composite scoring formula (traffic × demand × SVM) ^(1/3). However, the GA overcomes this limitation because it evaluates sets of stations together rather than one at a time. This revealed how important evolutionary optimization is for problems where objectives (like demand and traffic) may conflict.

3.5.3. District Coverage

The K = 20 selected stations cover 20 of 21 seniūnijas almost (95.2% coverage). However, only one district has not received any stations. The complete list of stations and there Cross-ponding district is available in Appendix 7 and 8.

3.5.4. Selected Station Characteristics

Table 30: Top 10 Selected Stations by Priority

Priority	Cluster	Seniūnija	Suitability Score	SVM Prob.	Traffic Score	Dist. to Existing (m)
1	90	Žirmūnai	0.863	0.193	0.746	489
2	105	Senamiestis	0.746	0.201	0.873	646
3	127	Karoliniškės	0.688	0.202	0.820	716
4	183	Pašilaičiai	0.659	0.358	0.443	493
5	115	Žvėrynas	0.654	0.246	0.871	504
6	191	Šeškinė	0.618	0.271	0.719	745
7	186	Rasos	0.616	0.193	1.000	555
8	24	Paneriai	0.608	0.193	0.058	2,291
9	198	Justiniškės	0.568	0.293	0.233	375

10	63	Pilaitė	0.419	0.236	0.020	2,008

(Žirmūnai) ranked at first, receiving a suitability score of (0.863). Moreover, Rasos has received the maximum traffic score of 1.000. However, Paneriai has received the lowest scores as it is an underserved rural area. Its distance from existing chargers is 2,291m.

3.5.5. Average Metrics of Selected Stations

Table 31: Metrics

Metric	Average Value
Traffic score	0.417
Suitability score	0.353
SVM probability	0.255
Distance from an existing charging station	650 m

Based on the results, the average distance of 650 m exceeds the minimum constraint set for 300 meters, confirming equal spatial distribution without excessive redundancy.

3.6. Phased Deployment Strategy

Based on results acquired from GA ranking, the selected 20 station sites were divided into three phases based on their priority ranking by GA

Table 32: Phased Deployment Plan

Phase	Stations	Priority Range	Avg Suitability	Avg Traffic	Avg Distance (m)
1	7	1-7	0.656	0.723	704
2	7	8-14	0.315	0.256	865
3	6	15-20	0.052	0.197	1,019

The station selected in phase one has the highest average suitability ranking of 0.656 with a traffic score of 0.723, suggesting immediate deployment. However, the selected station in Phase three has the lowest average suitability, but included to address the spatial coverage gap of underserved districts.

3.6.1. Visualization Outputs

The map (evcs_deployment_map.html) was developed, which includes proposed stations with their suggested phases (colored by phase), existing chargers (gray), suitable sites selected by SVM (red crosses), a suitability heatmap, traffic congestion hotspots, visible boundaries of seniūnija, and 400m areas around the proposed selection station site.

Overall Deployment Summary:

Table 33: Deployment summary

Metric	Value
Proposed stations sites	20
Covered seniūnijas	20
Suitability score (Average)	0.353

SVM probability score (Average)	0.255
Minimum distance to existing	301 m

3.6.2. Summary of Key Findings

Table 34: Overall summary of key findings

Component	Key Finding
Population	636,023 residents; 57.8% working age
Traffic congestion	Rasos highest (1.000)
Candidate zones	201 zones identified
Feature importance	POI density dominates (91.9%)
Ensemble performance	R^2 = 0.96 (internal consistency of proxy)
SVM classification	F1 = 0.84, Accuracy = 0.87
GA coverage	20 stations, 20 districts (95%)
Phased deployment	Phase 1: 7 stations,Phase 2: 7 stations,Phase 3: 6 stations

CONCOLUSION

By integrating multiple datasets, including population data (Vilnius total population of 636,023 residents across 21 districts), traffic congestion data from Vilnius open data portal, and POI data from OSM with 1.9 million POI across Vilnius into a unified spatial model, 201 candidate zones were identified.

In the proposed pipeline, the ensemble model (Ridge + Random Forest) has achieved R^2 = 0.96 on the constructed proxy; on the other hand, the SVM has achieved F1 = 0.84 in site classification of suitable locations based on revealed preferences.

Based on our targeted number of sites, the GA has successfully selected 20 stations covering 20 of 21 districts with a coverage of 95%, outperforming the greedy baseline by 0.3987 in fitness score.

A three-phase plan was produced for deployment, selecting the first 7 sites s as phase one, 7 for phase two, and 6 station sites for the third phase, respectively, along with an interactive map for stakeholder visualization and assessment. The full code of the implemented methodology and results can be found in Appendix 5.

Table 35: Comparison with Other Approaches

Approach	**Limitation**	**This Study's Advantage**
Distance-based heuristics	Urban activity patterns were not incorporated	Incorporate POI, traffic, and population data
Neural network demand prediction	It requires larger datasets, overfits (R^2 = -0.46)	Ensemble model performs good on small data (n=201)
Single-objective optimization	Cannot balance conflicting goals	Multi-objective GA with four weighted components
Unconstrained site selection	May place charging stations too close	Enforces 400m distance, 300m from existing charging stations

INNOVATIONS

This approach calculates true geometric areas for each cluster, improving spatial accuracy, unlike traditional circular radius assumptions. However, the ensemble method provides a continuous suitability ranking; Moreover, the SVM acts as an empirical feasibility filter based on revealed preferences of existing charging station operators. To avoid over-concentration, the fitness function explicitly penalizes districts and rewards geographic spread, ensuring district coverage of 95%.

RECOMMENDATIONS

For urban planners and policymakers:

Based on the obtained results, the first seven station sites 1-7 should be prioritized, having the highest suitability scores (avg 0.656) and traffic scores (avg 0.723). The map visualization will allow toggling of layers, clicking markers for details, and measuring distances. The current configuration suitability 35%, traffic 35%, population 15%, and 15% SVM prioritizes demand.

For more equitable coverage, increase SVM weight to 25% and reduce suitability to 25%. The proposed study provides a relative ranking. For absolute validation, actual charging session utilization data is required.

FUTURE WORK

The following directions are recommended for future research:

Suitability rankings should be validated against actual utilization EV charging session data when it becomes available. Incorporate time-of-use patterns and day-of-week variations in demand estimation. The GA should be extended to include installation and operational costs as additional optimization objectives. Test alternative weight configurations (e.g., equity-prioritized, cost-prioritized) for different policy contexts. Incorporate live traffic and parking occupancy data for dynamic recommendation updates.

LIMMITATIONS

The most significant limitation includes the absence of real EV charging session data. In the proposed methodology, the ensemble model predicts a constructed proxy, not actual demand. The reported values R^2 measure internal consistency, not predictive accuracy. All the obtained Results should be interpreted as a relative suitability ranking, not an absolute demand forecast.

The proposed framework, both the ensemble calibration (using distance correlation) and SVM labels (using distance thresholds), relies on existing charging station locations. The framework incorporates past charging station placement decisions, which may not be optimal in all cases.

64 zones obtained from the proposed methodology (23 presence, 41 absence) were available for SVM training, with 13 held out for testing. Results obtained may not fully capture urban diversity.

Several districts have shown zero traffic events, which may be because of a lack of monitoring rather than the actual absence of congestion.

REFERENCES

[Air24] Airlangga, G. (2024). Comparative Analysis of Machine Learning Models for Predicting Electric Vehicle Range. JURASIK (Jurnal Riset Sistem Informasi dan Teknik Informatika), 9(1).Link, alternative link

[AWF+25a] Ameer, H., Wang, Y., Fan, X., & Chen, Z. (2025). Hybrid optimization of EV charging station placement and pricing using Bender's decomposition and NSGA-II algorithm. Applied Energy, 397, 126385. Link, Alternative link

[Deb21a] Deb, S. (2021). Machine Learning for Solving Charging Infrastructure Planning Problems: A Comprehensive Review. Energies, 14(23), 7833. Link, Alternative link

[EA24a] Ezuruka, E. O., & Anigbogu, S. O. (2024). Developing an Efficient Support Vector Machine Learning Model for Mobile Charging Stations. International Journal of Research and Scientific Innovation (IJRSI), 11(8), 1492-1500.Link, Alternative link

[FFR25a] Ferraz, R. S. F., Ferraz, R. S. F., & Rueda-Medina, A. C. (2025). Optimized Planning for Electric Vehicle Charging Stations: A Systematic Review. IEEE Access, 13, 157059-157080. Link, Alternative Link

[HPD19a] Hahsler, M., Piekenbrock, M., & Doran, D. (2019). dbscan: Fast Density-Based Clustering with R. Journal of Statistical Software, 91(1), 1–30. (Note: The 2015 version you listed appears to be an earlier package release; the main associated paper is 2019.)Link, Alternative Link

[HW25a] Han, W., & Wang, D. (2025). New energy vehicle fast charging reservation algorithm based on internet of things coordination. Discover Internet of Things, 5, 124.Link, Alternative Link

[JPM+22a] Jordán, J., Palanca, J., Martí, P., & Julian, V. (2022). Electric vehicle charging stations emplacement using genetic algorithms and agent-based simulation. Expert Systems with Applications, 197, 116739.Link, Alternative Link

[KAK+26a] Khan, A., Al Hinai, M., Khattak, M. I., & Bait-Suwailam, M. M. (2026). Clustering-based intelligent electric vehicle charging station selection using dynamic user requests and station status. Sustainable Energy, Grids and Networks, 46, 102199.Link, Alternative Link

[WZW25a] Wang, Y., Zhao, J., & Wen, H. (2025). Multi-Objective Genetic Algorithm for Charging Station Capacity and Location Optimization. International Journal of Intelligent Information Technologies.Link, Alternative link

[MTG23a] Mehouachi, I., Trojette, M., & Grayaa, K. (2023). MONNA: Multi-objective neural network algorithm for the optimal location of electric vehicle charging infrastructure in Tunis city. Journal of Cleaner Production, 431, 139837.Link, Alternative Link

[PB23a] Palaniappan, A., & Bhukya, D. P. (2023). Data-Driven Analysis of EV Energy Prediction and Planning of EV Charging Infrastructure. In 2023 IEEE Ninth International Conference on Big Data Computing Service and Applications (BigDataService) (pp. 17-24).Link, Alternative Link

[PhD23a] Brunel University PhD Thesis. (2023). Optimization of Solar Home System-Electric Vehicle Charging Stations using NSGA-II and MOEA/D. Link, Alternative Link

[RYT+26a] Roozkhosh, F., Yao, X. A., Tahmasbi, B., Yang, H., & Mai, G. (2026). Bridging data gaps for EV demand estimation: A transfer learning approach. Journal of Transport Geography, 131, 104217.Link, Alternative link

[BOT+24a] Barakat, S., Osman, A. I., Tag-Eldin, E., Telba, A. A., Abdel Mageed, H. M., & Samy, M. M. (2024). Achieving green mobility: multi-objective optimization for sustainable electric vehicle charging. Energy Strategy Reviews, 53, 101351.link

[ZLW25a] Zhang, Y., Liu, X., & Wang, Z. (2025). Optimization of Electric Vehicle Charging Station Layout Based on Point of Interest Data and Location Entropy Evaluation. Sustainable Cities and Society, 112, 105847. Link, Alternative link

[ZZL22a] Zhou, G., Zhu, Z., & Luo, S. (2022). Location optimization of electric vehicle charging stations: Based on cost model and genetic algorithm. Energy, 247, 123437. Link, Alternative link

ABBREVIATIONS

ANN Artificial Neural Network

API Application Programming Interface

CSV Comma-Separated Values

DBSCAN Density-Based Spatial Clustering of Applications with Noise

EV Electric Vehicle

EVCS Electric Vehicle Charging Station

GIS Geographic Information System

IDW Inverse Distance Weighting

IoT Internet of Things

MAE Mean Absolute Error

ML Machine Learning

MSE Mean Squared Error

OSM OpenStreetMap

POI Point of Interest

R^2 R-squared

RBF Radial Basis Function

RF Random Forest

RMSE Root Mean Square Error

SVM Support Vector Machine

ε - Epsilon (DBSCAN neighborhood radius)

APPENDICES

Appendix 1. Vilnius population by age and district.

seniunija	population	young_share	working_share	senior_share	pop_25_64	pop_65_plus	
Naujininkai	30,129	0.2088	0.6341	0.157	19,105	4,731	
Paneriai	8,807	0.2651	0.6222	0.1128	5,479	993	
Lazdynai	23,932	0.2434	0.5692	0.1874	13,622	4,485	
Grigiškės	10,486	0.2356	0.5648	0.1996	5,923	2,093	
Vilkpėdė	19,310	0.2017	0.6009	0.1974	11,603	3,812	
Senamiestis	10,158	0.26	0.6227	0.1174	6,325	1,192	
Naujamiestis	20,857	0.2557	0.6082	0.1361	12,684	2,839	
Rasos	12,361	0.2598	0.5954	0.1448	7,360	1,790	
Karoliniškės	24,448	0.2182	0.536	0.2458	13,105	6,010	
Žvėrynas	12,089	0.3001	0.5485	0.1515	6,631	1,831	
Šnipiškės	14,652	0.2251	0.6277	0.1472	9,197	2,157	
Viršuliškės	4,014	0.2267	0.5424	0.2309	2,177	927	
Naujoji Vilnia	36,502	0.271	0.5831	0.146	21,283	5,328	
Šeškinė	31,102	0.2078	0.5527	0.2395	17,192	7,449	
Justiniškės	23,351	0.2363	0.5186	0.245	12,110	5,722	
Pilaitė	21,798	0.303	0.593	0.104	12,927	2,266	
Žirmūnai	43,453	0.2329	0.5702	0.1968	24,778	8,553	
Fabijoniškės	37,001	0.2531	0.5416	0.2053	20,039	7,596	
Pašilaičiai	40,384	0.3069	0.5579	0.1352	22,530	5,459	
Antakalnis	39,128	0.2848	0.5492	0.1661	21,489	6,498	
Verkiai	50,754	0.2772	0.5928	0.1299	30,088	6,595	

Contains a large set of data. Use the link to view the full table Link

Appendix 2: Full List of Candidate Zones.

cluster	lat	lon	poi_count	cluster_area_km2	seniunija	total_ev_score	ev_density	jam_score_norm	dist_to_existing_m	lag_traffic
67	54.68026	25.27492	1945	7.2579	Senamiestis	49.29	6.79	0.873	257.64	0.873
71	54.71419	25.29331	661	3.951	Žirmūnai	72.64	18.39	0.746	600.67	0.746
69	54.69819	25.28076	445	2.6012	Šnipiškės	53.67	20.63	0.69	386.92	0.69
2	54.70865	25.17962	329	2.9884	Pilaitė	17.91	5.99	0.02	320.79	0.02
75	54.70746	25.31443	193	1.1651	Antakalnis	15.03	12.9	0	127.38	0
80	54.69865	25.25549	168	0.9879	Žvėrynas	56.43	57.12	0.871	256.65	0.871
68	54.7246	25.23914	156	0.9819	Pašilaičiai	50.21	51.14	0.443	394.43	0.5
72	54.73776	25.22963	127	0.9439	Pašilaičiai	48.46	51.34	0.443	119.05	0.411
98	54.71535	25.24718	100	0.3496	Šeškinė	76.52	218.85	0.719	40.91	0.719
21	54.62313	25.14932	89	2.3072	Paneriai	4.52	1.96	0.058	2435.74	0.058
18	54.68815	25.41773	87	1.6798	Naujoji Vilnia	9.54	5.68	0	720.26	0
129	54.65131	25.2748	86	0.3157	Naujininkai	23.78	75.34	0	342.1	0
88	54.73923	25.2723	83	0.3093	Verkiai	24.88	80.45	0	241.01	0
4	54.66993	25.09617	67	0.5279	Grigiškės	18.34	34.73	0	2568.59	0
29	54.75436	25.27873	65	0.8591	Verkiai	11.03	12.84	0	143.42	0
30	54.76668	25.19711	61	1.5092	Pašilaičiai	41.69	27.62	0.443	3228.69	0.443
70	54.71047	25.26331	59	0.0822	Šeškinė	107.58	1309.22	0.719	93.57	0.709
17	54.75284	25.26349	51	0.1744	Verkiai	26.61	152.57	0	103.24	0
9	54.6522	25.30141	50	1.2437	Naujininkai	7.08	5.69	0	168.78	0

Contains a large set of data. Use the link to view the full table Link

Appendix 3: Feature importance ranking

Feature Name	Relative Weight / Importance
poi_density	0.9185308463
district_pop_density	0.02701650914
jam_score_norm	0.01886249385
lag_neighborhood_traffic	0.01457675187
lag_neighborhood_leisure	0.01186128546
lag_neighborhood_retail	0.006797029299
working_share	0.00235508404

Click the Link to the full file.

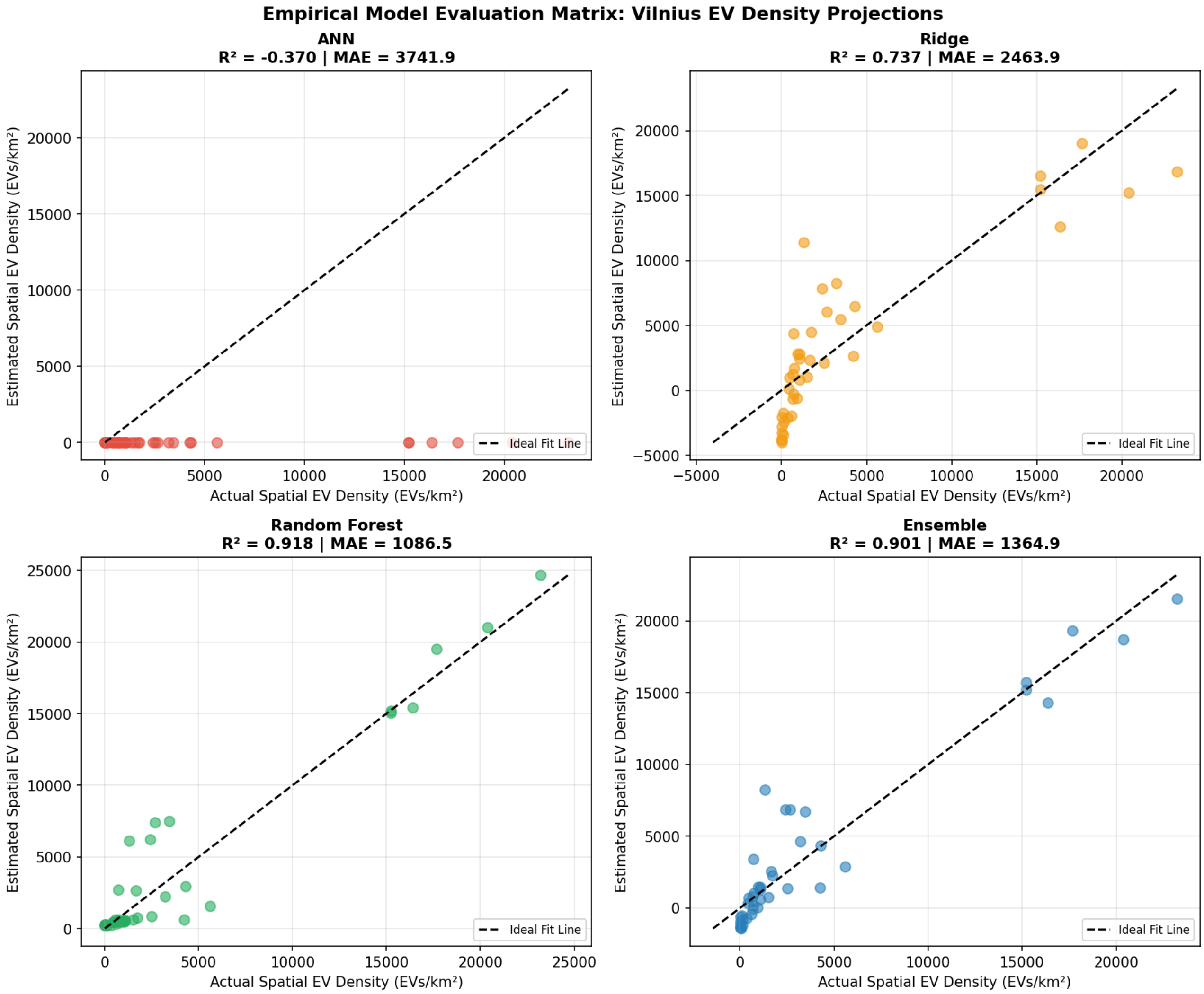
Appendix 4: Phase deployment map

(Download and Open in Chrome). Link

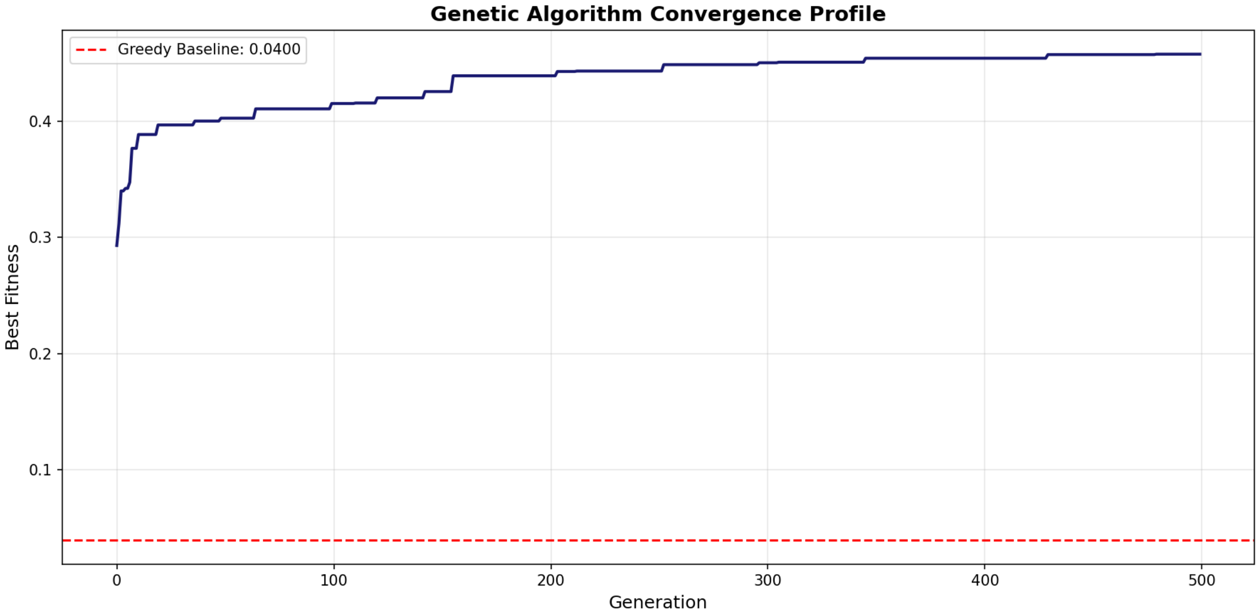
Appendix 5: Code Implementation

Full pipeline code can be found on the GitHub repository Link please read the README.md file for instruction

Appendix 6: Model Comparison



Appendix 7: GA Convergence



Appendix 8: GA Selected sites

The files is lengthy see the link for full file