Predicting Urban Greening Potentials with Artificial Intelligence Model

A GIS-based Machine Learning approach for local assessment

Wissenschaftliche Arbeit zur Erlangung des Grades

M.Sc. Umweltingenieurwesen

an der TUM School of Engineering and Design der Technischen Universität München.

Betreut von Dr.-Ing. Roland Reitberger

Lehrstuhl für energieeffizientes und nachhaltiges Planen und Bauen

Eingereicht von Vinayak Prem Kooniyara

vinayak.kooniyara@tum.de

Eingereicht am München, den 02.01.2025

Vereinbarung

zwischen

der Technischen Universität München, vertreten durch ihren Präsidenten,
Arcisstraße 21, 80333 München

hier handelnd der Lehrstuhl für Energieeffizientes und Nachhaltiges Planen und Bauen (Univ.-Prof. Dr.-Ing. W. Lang), Arcisstr. 21, 80333 München

– nachfolgend TUM –

und

Frau/Herrn Vinayak Prem Kooniyara

(Anschrift) Agnesstraße 31, Munich 80798

– nachfolgend Autorin/Autor –

Die Autorin / der Autor wünscht, dass die von ihr/ihm an der TUM erstellte Masterarbeit mit dem Titel

Predicting Urban Greening Potentials with Artificial Intelligence Model

A GIS-based Machine Learning approach for local assessment

☒ auf mediaTUM und der Webseite des Lehrstuhls für Energieeffizientes und Nachhaltiges Planen und Bauen mit dem Namen der Verfasserin / des Verfassers, dem Titel der Arbeit, den Betreuer:innen und dem Erscheinungsjahr genannt werden darf.

☐ in Bibliotheken der TUM, einschließlich mediaTUM und die Präsenzbibliothek des Lehrstuhls für Energieeffizientes und Nachhaltiges Planen und Bauen, Studierenden und Besucher:innen zugänglich gemacht und veröffentlicht werden darf. Dies schließt auch Inhalte von Abschlusspräsentationen ein.

☐ mit einem Sperrvermerk versehen und nicht an Dritte weitergegeben wird.

(Zutreffendes bitte ankreuzen)

Zu diesem Zweck überträgt die Autorin / der Autor der TUM zeitlich und örtlich unbefristet das nichtausschließliche Nutzungs- und Veröffentlichungsrecht an der Masterarbeit.

Die Autorin / der Autor versichert, dass sie/er alleinige(r) Inhaber(in) aller Rechte an der Masterarbeit ist und der weltweiten Veröffentlichung keine Rechte Dritter entgegenstehen, bspw. an Abbildungen, beschränkende Absprachen mit Verlagen, Arbeitgebern oder Unterstützern der Masterarbeit. Die Autorin / der Autor stellt die TUM und deren Beschäftigte insofern von Ansprüchen und Forderungen Dritter sowie den damit verbundenen Kosten frei.

Eine elektronische Fassung der Masterarbeit als pdf-Datei hat die Autorin / der Autor dieser Vereinbarung beigefügt. Die TUM ist berechtigt, ggf. notwendig werdende Konvertierungen der Datei in andere Formate vorzunehmen.

Vergütungen werden nicht gewährt.
Eine Verpflichtung der TUM zur Veröffentlichung für eine bestimmte Dauer besteht nicht.

Die Autorin / der Autor hat jederzeit das Recht, die mit dieser Vereinbarung eingeräumten Rechte schriftlich zu widerrufen. Die TUM wird die Veröffentlichung nach dem Widerruf in einer angemessenen Frist und auf etwaige Kosten der Autorin / des Autors rückgängig machen, soweit rechtlich und tatsächlich möglich und zumutbar.

Die TUM haftet nur für vorsätzlich oder grob fahrlässig verursachte Schäden. Im Falle grober Fahrlässigkeit ist die Haftung auf den vorhersehbaren Schaden begrenzt; für mittelbare Schäden, Folgeschäden sowie unbefugte nachträgliche Veränderungen der veröffentlichten Masterarbeit ist die Haftung bei grober Fahrlässigkeit ausgeschlossen.

Die vorstehenden Haftungsbeschränkungen gelten nicht für Verletzungen des Lebens, des Körpers oder der Gesundheit.

Meinungsverschiedenheiten im Zusammenhang mit dieser Vereinbarung bemühen sich die TUM und die Autorin / der Autor einvernehmlich zu klären. Auf diese Vereinbarung findet deutsches Recht unter Ausschluss kollisionsrechtlicher Regelungen Anwendung. Ausschließlicher Gerichtsstand ist München.

München, den	München , den 02.January.2025
……	……
(TUM)	(Autor:in)

Erklärung

Ich versichere hiermit, dass ich die von mir eingereichte Abschlussarbeit selbstständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

München, 02.01.2025 

Ort, Datum, Unterschrift

Table of Contents

Kurzfassung

Städte wachsen exponentiell und aktuellen Daten zufolge werden bis 2050 67 % der Weltbevölkerung in Städten leben. Städte werden oft als Hotspots für Treibhausgasemissionen und andere Anomalien verantwortlich gemacht, die Klimawandel und Umweltverschmutzung auslösen, aber wir erkennen nicht an, dass Städte durch effiziente und wirksame Planung und Politik auch eine zentrale Rolle bei Minderungsstrategien für diese spielen können. Nachhaltige Städte sind die Lösung und die städtische grüne Infrastruktur (UGI) ist eine der entscheidenden Komponenten. Die UGIs, insbesondere Bäume, erbringen viele Ökosystemdienstleistungen (ES) und ihre Entwicklung ist für die Nachhaltigkeit von Städten und städtischen Gebieten von entscheidender Bedeutung. Das Pflanzen von Bäumen in Städten und die Entwicklung von UGIs mit fortschrittlichen GIS-Technologien finden statt. Ihre Wirksamkeit bei der Erzielung realistischer Ergebnisse wird jedoch durch die unvorhergesehenen tatsächlichen Standortbedingungen und die mangelnde Datenverfügbarkeit beeinträchtigt. Weitere Herausforderungen bei GIS-basierten Studien sind der zeitaufwändige Wissenserwerb und die Notwendigkeit von Experten, mehrere Datensätze zu sammeln und zu analysieren. Um diesem Mangel entgegenzuwirken, kombiniert diese Studie die Potenziale der oberirdischen GIS-basierten Daten und maschinellen Lernmodelle, insbesondere Feedforward Neural Networks (FNNs), um diesen Prozess der Identifizierung von Standorten für Baumpflanzungen und Initiativen zur Begrünung von Städten zu automatisieren. Basierend auf dem neuesten Stand der Technik und Forschung wird eine Methodik mit mehreren Aufgabenteilungen entwickelt, die Datenaufbereitung mit Feature Engineering, Datenverarbeitung, Modelltraining und -test sowie Modellvalidierung umfasst. Zur Durchführung der verschiedenen Aufgaben der Methodik wurden mehrere Arbeitsumgebungen verwendet, wie z. B. ArcGIS Pro, IDE und Microsoft Office. In München wurde eine Fallstudie durchgeführt, die mehrere Datensätze umfasste, und um dieses Kunststück zu erreichen, wurden unterschiedliche Datenskalen verwendet. Das Framework, das verwendet wurde, um die Daten mit dem FNN-Modelltraining kompatibel zu machen, umfasste Feature Engineering, fortschrittliche Optimierungstechniken und iterative Prozesse, die die Leistung und Flexibilität des Modells verbesserten. Die Leistung des FNN-Modells wird mit der des Decision Tree-Regressionsmodells verglichen, und es stellte sich heraus, dass das FNN-Modell leistungsfähiger war. Das endgültige FNN-Modell erreichte eine hohe

Leistung mit einem R^2-Wert von 0,7916 und einem MSE von 228,68 und zeigte damit seine Fähigkeit, die Lerndaten zu verallgemeinern, um genaue Vorhersagen für die Zielvariable (Begrünungspotenzial) zu treffen. Die Vertrauenswürdigkeit und Transparenz der Modellvorhersagen werden mithilfe der SHAP-Funktionen und der Visualisierung der Vorhersagedaten in ArcGIS Pro interpretiert. SHAP-Werte könnten den Beitrag verschiedener Schlüsselmerkmale zu den Vorhersagen erklären, wie z. B. die Daten zur vorhandenen Vegetation, zu Oberflächenundurchlässigkeitswerten und zur Nähe zur nächsten Grünfläche, die mit den Beobachtungen anderer verwandter Forschungsarbeiten in ähnlicher Richtung übereinstimmten. Die iterative Improvisation und die Ergebnisse weisen auf die Flexibilität des Modells hin, das in anderen städtischen Regionen der Studie angewendet werden kann. Diese Forschung unterstreicht die Vorteile der Kombination von GIS und ML zur Optimierung der Ressourcennutzung und effizienten Entwicklung von UGI. Dieser Rahmen ermöglicht die Reproduzierbarkeit auf andere Städte, und dieses Modell kann Stadtplanern helfen, datengesteuerte Entscheidungen für die UGI-Entwicklung im Hinblick auf nachhaltige Entwicklungsziele zu treffen.

Abstract

Cities are growing at exponential rates, and based on current data, by 2050, 67% of the global population will be living in the city. Cities are often blamed as the hotspots for greenhouse gas emissions and other anomalies that trigger climate changes and environmental pollution, but we fail to acknowledge that cities can also play a central role in mitigation strategies for these through efficient and efficient planning and policies. Sustainable cities are the solution, and Urban green infrastructure (UGI) is one of the crucial components. The UGIs, especially trees, deliver many Ecosystem services(ES), and their development is vital to the sustainability of cities and Urban areas. Urban tree Planting and UGI development with advanced GIS technologies are happening. Still, their effectiveness in deriving realistic outcomes is hampered by the unforeseen actual site conditions and the lack of data availability. Other challenges in GIS-based studies are time-consuming knowledge acquisition and the need for experts to collect and analyze multiple data sets. To counter this shortcoming, this study combines the potentials of the GIS-based above-ground data and machine learning models, specifically Feedforward Neural Networks (FNNs), to automate this process of identifying tree planting locations and urban greening initiatives. Based on the state of the art and research, a methodology is devised with multiple divisions of tasks involving data preparation with feature engineering, data processing, model training and testing, and model validation. Multiple working environments were used to perform the various tasks of the methodology, such as ArcGIS Pro, IDE and Microsoft Office. A case study was done in Munich, incorporating multiple datasets, and different data scales were used to achieve this feat. The framework used to make data compatible with FNN model training involved feature engineering, advanced optimization techniques, and iterative processes that enhanced the model's performance and flexibility. The performance of the FNN model is compared with the Decision Tree regression model, and the FNN model was found to be better performing. The final FNN model achieved high performance with an R^2 score of 0.7916 and MSE of 228.68, exhibiting its ability to generalize the learning data to make accurate predictions on the target variable (Greening potential). The trustworthiness and transparency of the model predictions are interpreted using the SHAP functions and visualization of the prediction data in ArcGIS Pro. SHAP values could explain the contribution of various key features towards the predictions, such as the data on existing vegetation, surface imperviousness values, and proximity to the nearest green space, which were found to be aligned with the observations of other related research works done in similar directions. The iterative

improvisation and results indicate the flexibility of the model to be applied in other urban regions of the study. This research underscores the advantages of combining GIS and ML to optimize the resource utilization and efficient development of UGI. This framework allows reproducibility to other cities, and this model can help urban planners make data-driven decisions for UGI development towards sustainable development goals.

List of Abbreviation

AHP	Analytic Hierarchy Process
AI	Artificial Intelligence
Adam	Adaptive Moment Estimation
CNN	Convolutional Neural network
ES	Ecosystem Services
FNN	Feedforward Neural Network
GIS	Geographic Information System
IDE	Integrated Development Environment
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Error
NN	Neural Network
R^2 Score	Coefficient of Determination
ReLU	Rectified Linear Unit
RNN	Recurrent Neural Network
SHAP	SHapeley Additive exPlanations
UHI	Urban Heat Island
UGI	Urban Green Infrastructure

Glossary

Adam – An adaptive optimization algorithm is used in Neural Network machine learning models, which combines the advantage of optimizers like AdaFrad and RMSProp for efficient gradient-based optimization. [1]

Artificial Intelligence – Simulation of human intelligence in machines programmed to learn, think, reason, and solve problems. [2]

Batch Size – The number of training samples processed together for model training before the model updates its weights. [2]

Categorical features – Features used for representing discrete categories like land use types, street types, etc., which are often encoded using techniques like one-hot encoding to make features compatible with model training. [1]

Convolutional Neural Network – This is a type of Neural network used in image classification and object detection. It uses convolutional filters to extract features from the image files or grid-like data. [1]

Dropout – It's a regularization technique used in Neural networks to prevent overfitting where randomly selected neurons are ignored during training. [1]

Epoch- It refers to one complete pass through the entire training dataset during the training process of a machine learning model. [1]

Ensemble Model – A machine learning algorithm like the random forest that combines predictions from multiple models to improve the accuracy of the machine learning model.

Feedforward Neural Network – A type of Neural network used for regression and classification. They are computational models composed of multiple processing layers of Neurons to learn data representations with higher levels of abstraction. [2]

Geodatabase – A database used to store, manage and query spatial and non-spatial data. [3]

Geoprocessing Tools - These are software tools used to process and analyze geographic information, such as clipping the spatial data, buffer analysis, etc. [3]

Geographic Information Systems – A system for capturing, storing, querying, analyzing and managing geographic data like spatial data with descriptive information. [3]

Integrated Development Environment – A software suite like Google Colab that provides tools for scripting, code editing, debugger, and compiler for software development. [4]

Labelled data – Data used in supervised Machine learning tasks, where each input is associated with a known output. [1]

Loss Function – A function used in Machine Learning models to evaluate how well the model's prediction aligns with the actual values. [1]

Mean Square Error – A loss function used to measure the average squared difference between predicted and actual values in regression tasks in model training. [5]

Neuron – The fundamental unit of the Neural network, representing a single computational function. [2]

Normalization – Used in model training to transform or scale the input features to numerical ranges, which the model can learn. [1]

One-Hot encoding – A process by which categorical data is converted to numerical data to enable machine learning model training. [1]

Overfitting – A situation in which the machine learning model learns noises and details specific to the training set, which can result in the model performing well on training data but poorly on unseen data. [1]

Raster – Representation of spatial data in a grid-based form where each grid or cell holds the value or information of that data. [3]

Regression model – A machine learning model used to predict a continuous outcome (dependent variable) based on one or more input variables (independent variable). [1]

R^2 Score – A value ranging from 0 to 1 indicates the performance of the model, with a higher value indicating better performance. The coefficient of determination or R^2 Score is a statistical measure demonstrating the proportion of variance of the dependent variable explained by the independent variables in machine learning models. [1]

Rectified Linear Unit – ReLU is an activation function used in a neural network that introduces non-linearity and is computationally efficient. [2]

Shapefile – A vector data format used to store geometry and attributes of spatial features in GIS software. [3]

Spatial Data - Data containing locational or geographic components such as coordinates, boundaries, etc. [3]

TensorFlow - An open-source library supporting different model architectures for machine learning and deep learning applications developed by Google. [1]

Test Loss – A function used to measure the generalization performance of the model. It measures the quality of predictions of the model when evaluated on the test data. [2]

Training loss – A function used to measure the quality of the model's predictions in the training process when evaluated on training data. [2]

Underfitting - A situation in which the machine learning model fails to learn patterns and relationships on the training set, which can result in the model’s poor performance on both training data and unseen data. [1]

Validation loss- A function used to monitor overfitting during training, and it measures the quality of the predictions of the model when evaluated on the validation dataset.

1 Introduction

Cities have quantitative and sociological definitions, but in general, cities are relatively closed and dense developments with a concentration of workplaces, centres of trade, commerce, transportation, administrative and central functions, etc., which are socially, culturally, and economically productive. Cities vary depending on land use, location, country aspects, history of the city's evolution, resources, etc. The distinctive feature of all cities is their increasing number of inhabitants. [6]. Cities are growing at exponential rates, and based on current data, by 2050, 67% of the global population will be living in the city [7]. Most cities expanded and developed based on human needs, and many were developed through an economic lens of growth and demand [8]. The boundary limits of the cities are also increasing with increasing people and demands; for instance, it is projected that the Urban land in the United States is expected to increase from 3.1% in 2000 to 8.1% in 2050 [9, p. 65, 10]. Cities constitute a small area compared to the total land area, but their urbanisation process is felt globally in terms of climate changes, environmental problems, natural disasters, Urban heat island (UHI) effects in the cities, pollution, other hydro-meteorological and climatological hazards [11, 12]. The climate changes, heat stresses, and rising pollution in cities change the phenology, worsen the air quality, increase the energy and resource demands, deteriorate the health of inhabitants, and affect the natural biodiversity of the cities, leading to many ecological stresses [13, p. 147, 14–16]. The United Nation’s agenda for 2030, which include the 17 Sustainable Development Goals (SDG), has emphasised countering these adversities through SDG 3, SDG 7, SDG 11, SDG 12 and SDG 13 [17].

Cities are often blamed as the hotspots for Greenhouse gas emissions and other anomalies that trigger climate changes along with environmental pollution, but we fail to acknowledge that cities can also play a central role in mitigation strategies for these through good planning and good governance [18, 19]. With SDG 11, Sustainable cities are a solution to tackle multiple challenges while ensuring the needs of the inhabitants. Sustainable cities are human habitats that focus on the city's economic, social, and ecological aspects and the well-being of all inhabitants, so that they have the least possible impact on the environment. [20]. The misconception is that a sustainable city is to make cities look green, but in reality, it is a combination of green, blue and grey infrastructure with a permeable surface, spacious squares, canals, trees, hedges, green

grass parks, etc., for the well-being of the inhabitants [21]. Sustainable Urbanism is a holistic approach where an urban area is designed to enable Nature, and Urban Green infrastructure (UGI) is one of the crucial elements of sustainable urbanism as it forms the core part of enabling nature experience to human beings as well as a multitude of Ecosystem Services (ES) and promote biodiversity in the urban area[22, p. 4]. Making sustainable cities for its citizens requires UGIs, which are multiple functions oriented, an integral part of the urban environment, socially inclusive, and well-connected [20].

The UGIs, especially trees, deliver many ES, such as cooling by evapotranspiration, carbon sequestration, and shading, which can improve the microclimate around UGI in urban areas. UGI for a sustainable city are to be planned strategically, and their future development along with management is one of the critical aspects for successfully utilising the UGI to deliver ES. [23]. Many studies have quantitatively and qualitatively analysed trees' ES and validated the importance of urban trees. Trees in urban areas are found to reduce the surface air temperature and UHI, which positively impacts the comfort of inhabitants and reduces the demands on the resources utilised. Healthy urban trees in urban areas reduce the impervious cover around them, reducing the chances of flooding and summer heat from the impervious concrete surfaces. [24–26]. In the Urban areas, the places with higher tree cover experience better air temperatures than those without tree cover, which also enhances the biodiversity around the area. Trees play an essential role in improving the microclimate of the region near it, and studies have shown that trees positively impact communities and beautify neighbourhoods. Studies have found that trees are beneficial for the amelioration of air quality, reducing air pollutants like particulate matter, and the carbon sequestration process of Urban trees keeps a check on the rising carbon dioxide levels in the cities [18, 27, 28]. Psychological studies suggest that trees improve the quality of people around them and also the productivity of the lives around trees [26] [18] [29, p. 490, 30]. Apart from the various sustainability benefits and other environmental benefits, the significance of trees and UGI became apparent for the health of society during the COVID-19 pandemic [31]. Studies prove that UGIs are as crucial as the grey infrastructures of the cities.

Development of UGIs and planting new trees are becoming popular in cities, and many urban tree-planting programs are blooming in the cities and are becoming the need of the time [32]. In many countries, governments and other non-governmental organizations are pushing ambitious projects to plant more trees and develop the UGI in Urban areas, and one of the famous projects is Los Angeles' Million Trees LA

campaign [33, p. 2] [32]. Apart from tree planting projects, many UGI development programs are being implemented. As part of the development of UGI, Countries like Germany are preparing a system for effectively managing and developing their tree inventory using a tree information system. For this, very high spatial Resolution (VHR) Remote sensing techniques are used to gather data on urban trees [34]. The development of UGI and identifying tree planting locations with GIS studies further undertake these initiatives and lead to many advancements. Still, their effectiveness in deriving realistic outcomes is hampered by the unforeseen actual site conditions and the lack of data availability. [35–37]. Identifying tree planting locations is an elaborate process which requires the engagement of knowledge engineers, experts knowing multiple attributes related to the city or urban space, formulation of a sufficiently systematic approach, data on multiple facets of the above and below-ground features of the city and spatiotemporal datasets [26, 38–40]. Tree planting without a holistic view cannot solve the problem; 50 per cent of the newly planted trees are lost within 5 years after planting due to above and below-ground stressors for the tree growth [23, 40] [41]. Furthermore, urban trees have a higher mortality rate compared to natural forests. For instance, studies on urban trees in Baltimore in the US observed that the annual mortality rate of urban trees was 6.6% [41]. Hence, there is a need to improve the overall UGI development approach and find apt tree-planting locations for the expanding cities and urban lands.

This study aims to combine the potentials of the GIS-based data and Machine Learning (ML) models to overcome the lack of available datasets and reliance on time-consuming human efforts to identify the proper tree planting sites.

1.1 Aim of the Study

The study's primary objective is to improve the efficiency of UGI development with the help of GIS and ML. The development of UGI depends on multiple aspects of the city, and their scope for improving the city's liveability is vast. Location for new tree planting sites needs to address multiple questions such as: where are the areas most suitable for tree planting which are aligned with the goals of the organization who undertake the tree planting initiative, where are the locations which need tree planting, which areas are in critical to reap the maximum benefits of the ES from trees, where are the suitable location were trees can grow with proper nourishments and maintenance, and many other questions which are related to the provisions regarding the governing body of the

location [26]. Identifying new potential locations for greening within the city is a complex task requiring physical visiting of various sites, and there is a need to assess the site based on above-ground and below-ground data. With the availability of spatial and physical data on multiple features above the ground, the ML algorithms can understand the patterns and relationships between existing trees and other physical parameters of the city. These patterns and relationships between various city parameters can bypass the requirement of unavailable complex underground data. So, through this study, we are trying to combine the advantages of the available GIS-based above-ground data and the advantages of deep learning ML models to create a solution for the complex task of identifying the potential location for greening by making ML model to learn inherent patterns in the existing physical attributes of the city. Defining the target variables for locations suitable for planting trees in the program can be used to predict possible locations for growing new trees. So, this study intended to develop a framework for creating a dataset for a portion of the city to train ML models to predict the possible locations of trees in the other parts of the city.

We have taken Munich, Germany, as the case study location to accomplish the study. Munich has a good pool of GIS-compatible above-ground data about the various physical parameters of the city, which can be used to train ML models to learn the patterns and inherent relationships between various physical features.

1.2 Research Question and Hypothesis

Cities require more UGI, and their development is crucial for the sustainability of the cities and the delivery of ecosystem services. This study is intended to answer the research questions:

How well can the ML models trained with GIS-based above-ground data of a small portion of the city predict the greening potential of the other parts?

How can a framework be created to build a dataset that can help ML models learn the complex relationship between the physical and spatial components of the city?

How can the ML model predict areas with high greening potential in Munich?

The study aims to answer each research question by delineating the various processes defined in the methodology and their outcomes.

The primary outcomes expected from the research project are:

a) Converting the available GIS-based data to pure data, which can be used for model training.
b) Train the ML model to predict potential city greening locations using data from a portion of the city.
c) Comparing different model outputs and their validity.
d) Providing an overview of areas with high greening potential in Munich.

This study hypothesises that with the framework for the creation of data (which is limited to the above-ground features of the city) and training, the ML model can create an automated program which can be used to understand the complex relationship between various physical parameters of the city to identify target variables i.e., the greening potential as per this study. This methodology and AI ML model can be reconfigured according to the need and can be expected to be employed in cities with similar GIS data on the city's above-ground features.

1.3 Significance of the Study

Tree planting in cities is a comparatively new phenomenon that emerged as a resolution to the detrimental changes in cities owing to the rapid urbanization and industrial revolution, which affected the lives of city inhabitants [29, p. 477]. Planting new trees in cities requires permission from various authorities and knowledge-based analysis on multiple data such as the tree data, climate factors, soil characteristics, growing space, site condition and location, existing vegetation, land typology and its ownership, land regulations, social influences, street data, government regulations for urban development, sidewalk data, underground lines and other MEP data, other government data about the underground data [9, 35]. There are rules and regulations on one side, the city's underground lines and other essential service lines; the above-ground factors also play on the other. In many locations, data unavailability, such as the water line channels, electricity lines, mobility and other data features of city administration and national security. [40] [13]. From Figure 1, multiple stressors are mentioned that can interfere with the development and growth of the UGI, especially the trees in the city. The below-ground stressors include hindrance to the development of roots due to underground grey infrastructures, including utility lines of the city administration, poor soil composition caused by non-natural soil types, poor water infiltration, lack of nutrient availability in the urban soils, pollutants in the soil, and many other urban underground factors. The above-ground stressors include restrictions from buildings and other grey

infrastructures, surface sealing, vehicles, human interventions, pests, increased heat due to grey infrastructures, and many other restricting factors above the ground. [22, 40]. Hence, manually finding a location is a humungous task for finding new tree-planting sites within the city.

In cities with 80 to 90% of land covered by grey infrastructures, land availability is one of the controlling factors in identifying new tree planting locations, and it requires proper urban planning visions with an overall holistic view and knowledge in the development of UGI. Tree planting drives or programs are rarely utilized near the zones with a lot of Grey infrastructure due to a lack of data or understanding of the utilities and underground structures near the grey infrastructures. Hence, the sites proximal to the grey infrastructures are not utilized for their potential for UGI development. [22, p. 13]. Also, the expansion of the cities and the urbanisation process of the near regions of the cities will have to be under consideration of the UGI development drive as these regions will be experiencing a contraction in land availability and the detrimental effects of the urbanisation processes in the future. [22].

As a step up to improve the efficiency in the development of UGI, GIS-based methodologies have shown better management and planning of these developments. A combination of remote sensing and GIS data is considered an effective tool in data-driven decision-making and creating knowledge-based systems to aid the holistic development of the topography or Urban areas [38]. Many GIS-based studies on different cities were based on unique goals of UGI development; some studies were engaged in improving the microclimate and reducing UHI of the cities, and some others were focused on improving the water infiltration in Urban lands by focusing on tree planting locations in non-used urban lands, some concentrated the studies on enhancing the quality of the Urban habitat, other studies focused on identifying tree planting locations to reap maximum benefits of other ES from the trees, and some general GIS-based studies to identify the tree planting locations in cities with limited data [35] [22, p. 13] [42] [26, 39]. Lack of sufficient data was one factor affecting the GIS-based studies, and the necessary data for deep analysis to conduct need-based studies were challenging to access as this spectrum of data is spread across different organizations or departments. In the case of the accessed data, some may not be compatible or available in readily usable formats for GIS analysis by the experts. [35] [26]. The GIS-based approach requires experts to consider investing a lot of time and applying knowledge from different points of view and discipline, which can further delay the process [22]. The time-consuming and resource-intensive knowledge acquisition

program in deriving knowledge-based and data-driven systems is one of the major bottlenecks to tackle. [38].

The advent of ML, a subfield of Artificial Intelligence (AI), is the science of modelling computer programs for the human learning process by which automated knowledge acquisition is achieved. With ML, the existing data is used to train the AI model, which uses a rule-based system to enable the model to acquire knowledge and theories using inference strategies to make predictions. While many studies have integrated GIS and ML in the urban context, there has been no approach yet to determine an automated model to identify new tree-planting locations.

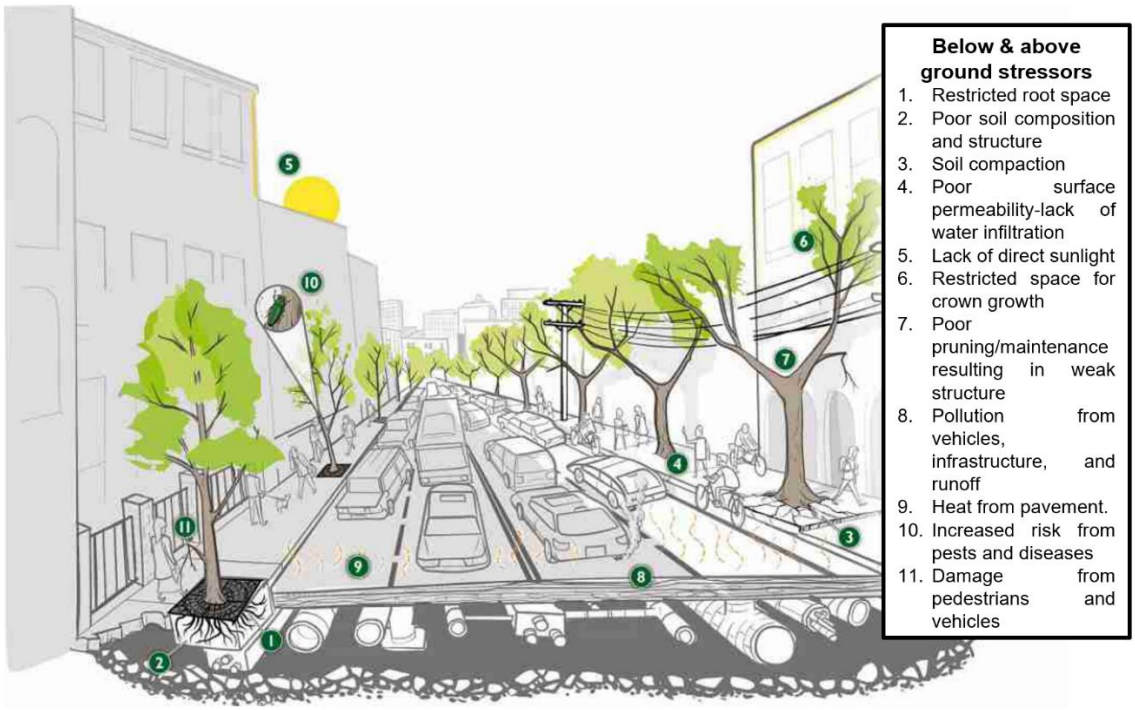

Figure 1 | Below and above ground ground stressors for Urban trees [40, p. 316]

2 State of Art

The expansion of cities and urbanization processes are reshaping the topography and landscapes worldwide, which triggers multiple challenges in the course of the process. Increased grey Infrastructures and urbanization lead to reduced green spaces, increased impervious surfaces, misalignment from the land's natural topography, UHI and other environmental issues in the Urban areas. As a measure to handle these adversities, Urban planners are increasingly focussing on data-driven systems for informed decision-making to assess and develop the UGI. Geographical Information system (GIS) - based UGI development is now widely used to manage and plan sustainable cities, requiring manual identifications and multiple analyses. [43–45]. With the development and progress in interoperability among different software and data formats [46], there is a huge possibility of data exchange between various platforms and GIS [47]. The vast pool of data from remote sensing and other GIS-compatible data of the cities requires data management and processing at various levels [45, 48, 49]. GIS-based ML researches are widely employed in urban and regional contexts to assess the region's spatial features and physical parameters [50]. With ML, these data can be used to process various functions such as filtering, interpretation and predictions. The availability of multiple ML model types for different applications is another key aspect of these models, which interpret relationships and patterns in the distribution of various physical parameters of the city for making predictions with the available data. [49] . This study focuses on utilising the advantages of ML and a dataset derived from GIS to develop UGI in the cities. This section reviews the methodologies relevant to urban tree planting, ML and GIS-based UGI development studies, and the works that can aid the study.

2.1 Urban Tree planting and Spatial parameters in urban space

Urban tree planting requires a comprehensive understanding of the urban landscape's various spatial and physical parameters, as these parameters can significantly influence the selection of new tree planting sites in Urban areas. The various parameters chosen for the study are based on the researchers' propositions regarding the factors considered for tree planting sites in urban areas. Multiple criteria need to be considered

for Urban tree planting, and the primary spatial and physical parameters to be considered are climatic data, site data, existing vegetation, land regulations, land utility data, social influences, maintenance requirements, grey- infrastructure data, accessibility to the site, soil data, underground infrastructures data. [9, 35, 51]. There have been research studies that are done to devise strategies for urban tree planting, and the main three categories are: (1) good physical growth of the tree and ensuring the survival rate of the planted trees, (2) maximising the Ecosystem services from trees and environmental benefits, (3) enhancing the aesthetics and liveability [35, 52, 53]. Studies suggest that combining these parameters using the GIS tools can provide a systematic approach for locating new tree planting locations and improving the greening potential of the Urban area [35].

Studies have been done to devise measures to ensure the development of UGI and suggestions to identify new tree-planting locations in urban areas. Studies done by Rieke Hansen & et al.[22, p. 14] has formulated five basic principles for the development of UGI in urban areas, which are: improving the quality of the UGI, promoting the diversity of functions and multiple uses of UGI, Connecting the green areas and creating green networks, encouraging cooperative alliances and endeavours, developing grey infrastructure and UGI in tandem. Multiple research suggests goal-oriented tree planting and UGI development measures; for instance, locations with higher levels of air pollutants, a higher density of inhabitants, and high traffic volume areas can be ideal locations for planting new trees to reap maximum ES. [26, 42]. Planting more trees near existing habitats, impervious covers, and comparatively higher temperature areas than the adjacent areas can help reap maximum environmental benefits and ES from trees. For instance, In Germany, there is a need for upgrading the drainage systems and transportation facilities, and these developments of grey infrastructure can used as opportunities to redesign these grey infrastructures in collaboration with UGI. [22, 26]. Research has been conducted to formulate a systemic approach for tree planting initiatives, which suggests connecting the desired benefits of trees like ES with the objectives, targets, and indicators. In this case, an indicator is delineated as the canopy cover percentage or crown projection area of trees. [54, p. 2].

GIS-based urban and environmental studies have incorporated data on the spatial and physical features of the above-ground attributes of the urban area created through remote sensing and photogrammetry. These spatial and physical data are used in tandem with GIS tools to derive useful information that can be put together to develop methodologies to improve urban land use patterns and urban green spaces. [55, 56]. A

study done on prioritizing the tree planting locations on streets used spatial and temporal data like surface temperature, existing tree density and vegetation cover, tree functional diversity, land use types, street data, socioeconomic data, and tree data to help the planners make decisive locations to promote tree planting along the streets [18, 32].

2.2 Framework for GIS-based Machine Learning Models

City space constraints play a significant hurdle in locating new tree planting sites and improving the greening potential of urban areas, as available spaces are underutilized or obscured due to a lack of knowledge. Identifying new tree planting locations requires extensive data collection on various spatial parameters and strategic data. Using ML to predict the greening potential of a city or urban area by training the model with the holistic data of the city is a complex task involving multiple processes. The framework for creating the dataset involves joining numerous data of the city's various physical and spatial parameters. In this study, the greening potential in the context of trees is considered for the UGI study since the trees have longer life spans and form a considerable part of the UGI [57, 58]. While there are studies incorporating GIS for identifying new tree planting locations, the major bottlenecks these studies face are the limited availability of data, physical site validation, and different goal orientations of studies, resulting in the absence of standardized methodologies for achieving the study's objective. Many studies were utilized in similar fields to employ GIS-based data to find optimal locations for various urban-based features. [35, 59].

There has been research on finding optimal sites for placing solar panels, which utilize spatial parameters defined by attributes, physical dimensions, position, extent, and spatial consistency, which are processed for multiple criteria analysis using an Analytic hierarchy process [60]. Mark C. Dwyer and Robert W. Miller used GIS to assess the benefits of urban tree canopies and surrounding land use by evaluating the environmental benefits of urban tree canopies, such as cooling effects and carbon sequestration. This study highlights the importance of GIS in understanding the interaction of urban tree canopy with the surroundings. [61]. GIS-based approaches were used in identifying ideal tree planting locations along the streets, which incorporated multiple spatial and temporal data to define the weights for the indicators to decide the locations to be planted with new trees [18, 32]. Studies done by Dexter &

et al. [27] prioritized tree planting sites based on goals such as meeting the community's needs and the location's suitability. This study has created the need-based criteria and suitability criteria by agglomerating various features coming under each criterion, and these variables are assessed in GIS tools to identify the new tree planting locations. As part of the MillionTrees initiative in New your City, a study was done to identify the best tree planting locations to reduce air pollution by prioritising GIS tools to filter the three indicators, such as pollution concentration, population density and low canopy cover, to make them the basis for identifying new tree planting locations. The paper introduces GIS methodologies that prioritise indicators such as urban tree canopy enhancement as a strategy to identify tree planting sites. [62]. There have been multiple GIS-based approaches for urban-related studies and identifying new tree planting locations, but their objectives and goals differ. There has been no standard methodology for a holistic approach to identifying new greening locations within the city based on available data; even if it is present, the limited data availability and validation studies are shortcomings of the studies [35].

The study by C. Wua & et al. identifying tree-planting locations using GIS relied mainly on the remote sensing data, which analysed the tree sites based on their land cover, permeability conditions of the land, existing vegetation, and type of land. The initial phase of the GIS framework was to identify the locations with potential planting sites using the remote sensing data. In the second phase, iterative modelling processes were used to determine the number and type of trees that could fit into these newly identified locations for tree planning. The research had limitations on the accuracy of the data and the unavailability of other significant data sets such as the street data, building data and other physical parameters of the urban area. [35]

Urban tree placement Analysis study done by R. Reitberger & et al. underscores the effectiveness of GIS analysis on above-ground spatial parameters like land cover, surface imperviousness data, existing vegetation, and urban infrastructure in identifying tree planting sites. This study has emphasized optimizing the UGI development strategies through spatial analysis, data processing, and selecting features that can directly affect the outcome. This study also suggests strategies for formulating the feature engineering of data to enhance the study's performance and effectiveness. [39].

C. V. Ekeanyanwu & et al. [63] reviewed the methodologies for merging the GIS and ML to find efficient data transfer methods to reduce loss and adaptability of GIS-based ML models for productive predictions and automation. Joan M. Peralta and Thelma D. Palaoag explored the possibility of using ML from GIS data to map Urban Green space

in smart cities, and the model used for this study was found to perform with high accuracy for predicting the data to help urban planners with data-driven decision making [64]. Similarly, the ML models were used in studies to assess and analyse the quality of green spaces, and the performance evaluated for the ML models suggests the model for practical implementation and usage due to their high accuracy [65].

2.3 Machine Learning Models

ML, an application of AI, is now widely used to make computer programs automatically learn data and improve based on the learning processes [66]. Multiple ML algorithms are used based on the purpose and datasets. ML algorithms like supervised ML algorithms, unsupervised ML algorithms, semi-supervised ML algorithms, and reinforcement learning ML algorithms are the algorithms that are used the most. This study focuses on the open-source library for ML frameworks like TensorFlow using Keras, which is a high-level API designed for easy model building [67]. TensorFlow is flexible and uses an extensive library of models, which provides a comprehensive ecosystem for building and training ML models [68].

2.3.1 Supervised Machine Learning Algorithms

The supervised ML algorithms are used in the GIS-based ML programmes due to their proficiency in learning labelled data to predict the target variable in unseen data based on the learned data. [69] .Supervised ML algorithms, like ensemble models, are found to be suitable for handling labelled spatial data to map input variables to desired outputs and predict urban development patterns in the case study of North Carolina [5]. The supervised ML algorithms have more flexibility in terms of applicability in diverse datasets and scales, making them compatible with handling extensive spatial data from GIS-based studies. Studies on GIS-based ML programs show that supervised algorithms provide insights into the features that can aid the interpretability of the model. Studies focussed on urban mapping analysis using ML suggest that supervised ML algorithms are designed to reduce errors in prediction by real-time monitoring of the progress of the models' training. From the study on urban mapping analysis using multiple ML algorithms, the Neural Network (NN) ML algorithms performed better and had high accuracy in predicting data involving remote sensing multispectral satellite imagery datatypes. [64, 70–72]. Supervised ML algorithms demonstrated higher efficiency and interpretability in the GIS-based ML studies used in Urban land planning

and environmental assessments. The performance of novel supervised ML models effectively dealt with complex GIS data validated in the studies related to the Urban growth predictions of Nasiriah City in Southern Iraq [71]. Supervised ML algorithms like NN can learn non-linear relationships in large-scale datasets derived from environmental and urban contexts and performed better [73][55]. The decision tree models, Ensemble models and the NN model types in TensorFlow are the major model types used to deal with spatial data and multiple decision-making conditions [74]. The model type to be chosen is decided based on the dataset type, the dataset's complexity, scalability, and the requirement/objective of the model prediction [67, 74].

2.3.2 Data Types and Feature Engineering

The input data for model training plays a pivotal role in the performance of the ML model. The GIS-based data encompasses different types of data, predominantly structured, unstructured and spatial. The structured data comprises well-defined, organized data stored in rows and columns, usually in Excel formats or SQL databases. The unstructured data comprises free-form data that does not follow any specific order or division of information in images, pixels, etc. [1, 2]. Remote sensing imagery development-related studies with ML utilize unstructured datasets to train the model, but too many layers of images or complexities can hinder the performance of the ML models on unstructured datasets [75]. Spatial datasets are found in file formats like GPS coordinates and shape files used in geospatial analysis, containing data that are geographically referenced [76]. Different datatypes are used for different purposes of model training, and the scale of the data availability also interferes with model choice. The choice of the data used for ML model training depends on the source of the data, the formats of the data, the objective of the study and the complexities of the target variable, which is labelled for making predictions. [2].

Creating a dataset for training ML models which can be used for predicting labelled data of complex target variables like the greening potential or new Urban tree planting locations within the city requires information from multiple GIS data of the urban area ranging from raster files to the point data features [39, 76]. Multiple data types in GIS encompass their kinds of values, information and patterns. GIS data of urban-based spatial and environmental data is a mixture of multiple data with different sub-regions, which add even more complex patterns in the dataset generation. The scale of the dataset is another decisive factor in GIS-based ML training, as some ML algorithms cannot perform with varying and very large datasets (which are common in urban-

related studies involving spatial data). [2, 77, 78]. Preprocessing the dataset to handle missing data and other noises in the dataset for ML model training and relative ease in achieving the due process also plays a significant role in the choice of the dataset used for ML algorithms. Studies show that geospatial and spatially heterogeneous data faces preprocessing challenges and their influence on ML model selection. The unstructured data requires specialized pre-processing and transformations to make it compatible with ML model training. In contrast, structured data is advantageous on those fronts because of its ease of handling. [77, 79]. Handling and extracting datasets for labelled ML model training involves multiple data features and pre-processing.

Feature Engineering is the process of transforming raw data into meaningful features and creating spatial relationships between the labelled target variables and the other training features of the ML model, which can improve the model's performance. Spatial feature engineering involves creating new variables or features and spatial relationships within data, enabling better ML model performance by learning patterns and relationships. A study involving urban vegetation mapping uses spatial feature engineering processes to train the NN models to enhance their performance. [80]. A study done on the training NN models in the Digital Elevation Model (DEM) height estimation of different terrain types utilizes the experimental data to be feature-engineered to make the raw data compatible for the model to effectively learn patterns and improve the model's performance [81]. Handling multiple data features and combining different data types for use in the model to learn patterns require effective feature engineering to enhance the performance of the ML model, as it helps reduce the overfitting of the learned data for the model. For instance, Feature engineering is essential for transforming categorical data to numerical representation and normalising data to values 0 and 1, as ML models learn on such data. Similar feature engineering operations were used in the study to optimize the performance of the ML model, which aimed at measuring urban green space exposure based on street view images. [4, 82]. Feature engineering in Urban studies is crucial for transforming data and multiple operations like normalization of data, categorical encoding, and spatial feature creation. [2]. Studies done on NN models on spatial data have found that custom feature engineering can enhance the predictive ability of the model [83]. A NN ML study on Gully erosion susceptibility has derived independent variables like distance to a river, drainage density and 11 other features through feature engineering to improve the efficiency of the NN model [73].

Considering the requirement of the data types, preprocessing, handling the data for easy manipulations, feature engineering, and source of data, the structured dataset is the most practical and easy option for handling multiple features [79]. Structured data are utilized in regression models where the source of the data was from GIS-based studies [84]. The structured data can also include spatial attributes and complex feature extraction. A study by Nikparvar & Thill found the handling and simplicity of structural data for NN ML models for handling data encompassing spatially heterogeneous data [77].

2.3.3 Neural Network Models

NN are ML models which are inspired from the Human brain. NNs are made of layers of interconnected neurons which can learn patterns and relationships to represent data through model training. Multiple types of neural network models are designed for defined purposes. The main Neural Network models are Feedforward Neural Networks (FNN), used for general-purpose tasks with structured data; Convolutional Neural Networks (CNN), used in image classification and object detection; and Recurrent Neural networks, used in Natural Language Processing (NLP). [2, 67]. NN models can capture non-linear dependencies in the data, the scalability in NN is standardised, and they can handle heterogeneous large datasets. [1]. NNs are widely used in Geospatial and remote sensing fields due to their ability to combine multiple features and heterogeneous geospatial data [71, 85]. NN models perform better in studies focussed on Urban growth predictions and land use classifications due to their interpretability of the training progress and flexibility in handling data [82, 86]. In the research that focused on urban green space mapping in smart cities, the NN model demonstrated higher accuracy than other models like Support Vector Machines and Random Forests [64].

FNN, also known as Multi-layer Perceptrons (MLP), is found to be particularly effective in performance in structured tabular data and has shown significant validity in modelling spatial data for their ability to learn non-linear relationships, which exist in Urban landscape studies [87]. FNNs are computational models composed of multiple processing layers of Neurons to learn data representations with higher levels of abstraction [88]. Studies on FNN and urban land use suggest that the FNN algorithms can learn non-linear interactions of diverse feature types like numerical, categorical, and spatial data [89]. The FNN has been applied in urban studies, which utilize structured datasets to predict urban traffic, incorporating multiple features like road networks, vehicle counts, and temporal data [90]. FNN models were found to perform with high

accuracy in predicting urban building energy consumption and urban road land use planning, and they were trained for structured data containing spatial data [56, 91]. Studies on GIS-based data and FNN with structured feature-engineered spatial data demonstrated better prediction on spatial data with improved feature engineering of training datasets [73].

2.3.4 SHAP Values for Model Interpretability

Training models with complex heterogenous data makes the model more complex and their interpretability is critical for their prediction explanation, debugging and data exploration. Shapely Additive Explanations (SHAP) values give insight into quantifying the various features contributing to the model's prediction. [92]. In GIS-based ML training program studies involving multiple features and large datasets, SHAP values help explain the relative importance of the various features used for training the model that can positively or negatively affect the predictions. [93]. The SHAP method is found to be seamlessly integrated with supervised ML models and NN, thereby enhancing the trustworthiness and transparency of the model's predictions. The SHAP values are visualized using the Beeswarm plots, Waterfall plots and other types of plots. The Beeswarm plots give the global overview of all the SHAP values of the features selected for training, and the waterfall plot provides the SHAP values of a single sample prediction with data on various features affecting that prediction. [94]. The use of SHAP in the context of Urban studies with ML models can provide actionable insights and help interpret the ML model's predictions and methods to improve the model's performance by understanding the dependencies of various features in the dataset that affect the prediction. [95].

3. Methodology

The overview of the research is to create a decisive framework to convert the spatial and physical data of the city in the GIS tool to data formats, which can be used to train the AI models to identify the patterns and relationships between the UGI and above-ground data.

The methodology was derived based on thorough review of the state-of-art approaches and research. The methodology of this research is subdivided into multiple tasks/stages of progress to achieve the desired goal. The skeletal flowchart of the workflow or stages of the methodology is represented in Figure 2.

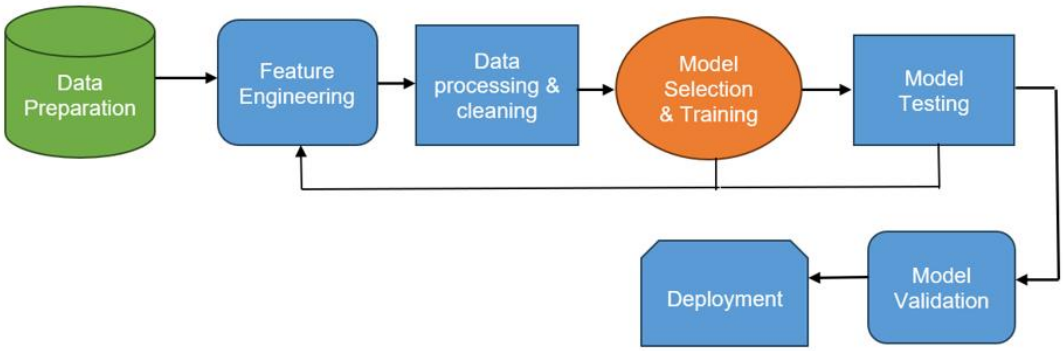

Figure 2 | Skeletal flowchart of the tasks in the methodology

Tools Used

ESRI ArcGIS Pro version 3.4.0 (ArcGIS) is used for the GIS part of the study. Analysis and geospatial processing of various data gathered are carried out in ArcGIS. Integrated Development Environment (IDE) Google Colab is used for Python scripting the AI model with packages from tensorFLow (2.12.0) and the SHAP functions. Microsoft Office was utilized for data processing and clearing the noises in the data.

3.1 Data Preparation

Data preparation is the methodology's initial phase, which involves collecting and preparing the collected data in the GIS platform (Here ESRI ArcGIS Pro). Based on the insights from the state-of-the-art section, all the available and relevant data are added to the working map environment of ArcGIS from the “catalogue pane” of the ArcGIS

working map contents pane. The data used are in the formats Geodatabase files, shapefiles, Raster data, and GeoJSON formats [96]. The data not readily available in shapefiles or formats supported by ArcGIS are imported into ArcGIS using the “Quick Import” data interoperability tool in the Geoprocessing tools. After importing and adding the shapefiles into the ArcGIS, all the data layers are checked for the spatial references and projected coordinate system. The projected coordinate system used for this study is ETRS 1989 UTM 32N. All the data added are projected into the same Coordinate system using the “project” tool in ArcGIS.All the shape files containing the detailed physical and spatial data on various physical parameters of the city, such as data related to trees, buildings, land types, street data, and perviousness of the land, are then successfully overlaid and checked for accuracy and alignment with the base maps available in ArcGIS. Geoprocessing tools, like “Raster to Polygon”, are used to convert the Raster files into shape files to enable straightforward data derivation. Figure 3 shows the ArcGIS map illustration of the data preparation task used in the case study of Munich with the integration of various data used in the study.

The data is then clipped using the “Clip” tool in ArcGIS according to the extent of the area considered for the respective studies, such as the pilot study area, main study area, and validation study area, which will be regarded in the later sections.

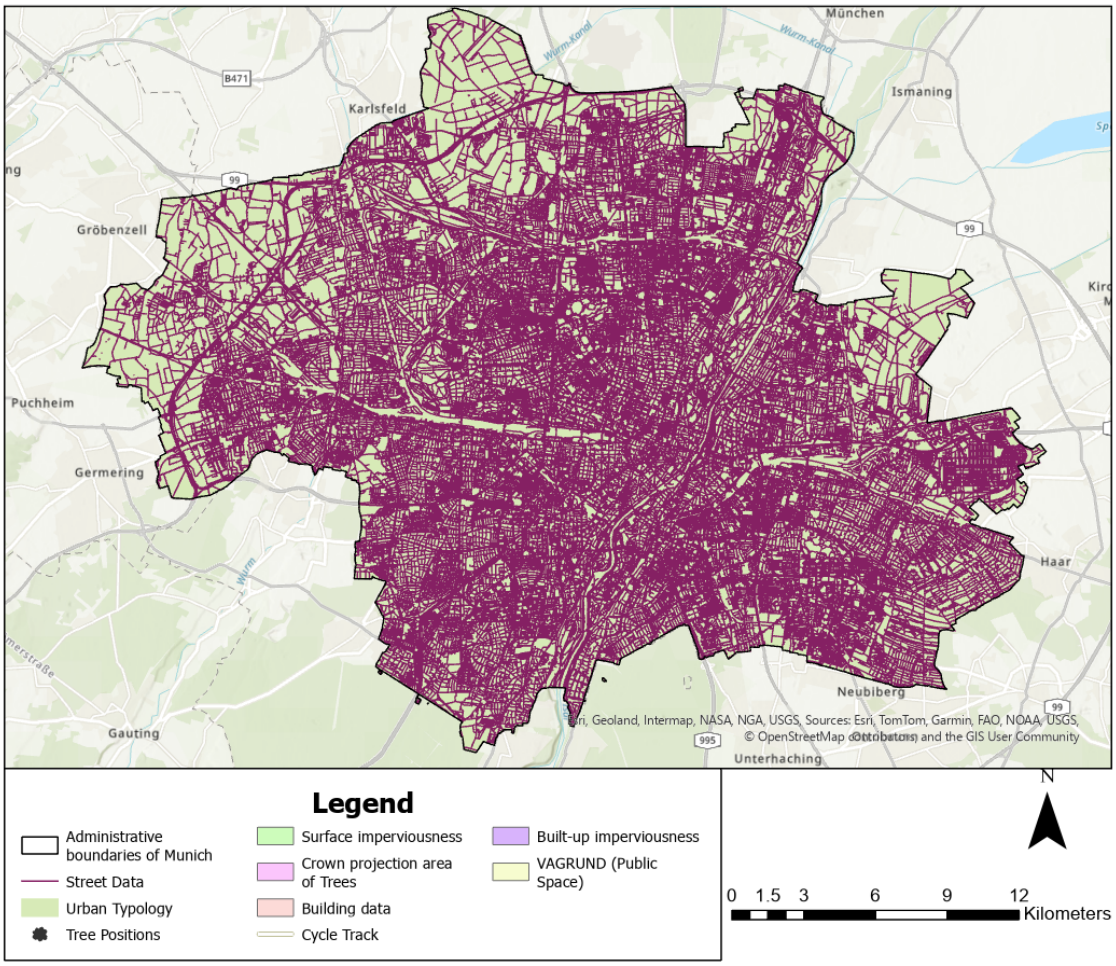

Figure 3 | GIS illustration of data preparation task used in the case study of Munich

3.2 Feature Engineering

As outlined in the state-of-the-art review, feature engineering is done to create a relationship between various physical features of the city so that these inherent complex relationships between the different physical attributes of the city can be used to train the ML model. Feature engineering involves creating a framework for a hierarchical process in the ML model training. This consists of defining the target variables for training the ML model and creating physical relationships between various available physical attributes of the City used for the study. Furthermore, deriving additional features for the model to learn the inherent interaction of various above-ground physical features such as complex interactions like the building and tree density within a certain study radius, land use patterns at different urban zones and other dependencies. The dataset is created by bringing the relationship between the target variable and other data features into the purview of the model training.

For this study, we have derived a process through which the model will learn the relationships between the city's various physical and spatial features. The area chosen for study is first defined according to the extent of the area of the study (Pilot Study, main Study, validation study). The structured tabular data was used in this study to train the model. The relationship of the city's various physical and spatial parameters was defined with respect to a reference point. For this approach, a new set of points was to be introduced in the study area.

In ArcGIS, the study area is first defined, and all the data used for the study is clipped to the extent of the study area. A square grid of 10m in size is introduced in this study area using the tool “Generate Tesselation”. The attribute table of the square grid is updated by adding two fields: centroid-x and centroid-y. The centroid of each square grid is then calculated by the tool “ Calculate geometry attributes”. To create a point feature in these centroids of the square grid, a tool called the” XY table to point” tool was run to introduce points using this centroid-x, centroid-y values of the square grid feature. This point in the centroid of each square grid is used as the reference point from which the relationship of various physical features of the city is defined, and the square grid is considered the area of influence for these grid points. The near distance to trees, streets, and buildings is derived using the near tool command. With these newly introduced grid points as reference points, the distance to various target features is calculated using the “Near“ tool.

- Distance to the nearest trees from the grid points, with a search radius of 7.05m (Search radius is considered in the purview that half the distance of the diagonal of the square grid of 10m size is 7.05m)
- Distance to the nearest building from the grid points, with a search radius of 5m (Search radius is fixed based on the results from running multiple iterations of model training and the influence of this particular feature in model training using the SHAP values).
- Distance to the nearest street from the grid points, with a search radius of 7.05m.
- Distance to the nearest cycle track from the grid points, with a search radius of 7.05m.
- Distance to the nearest “open or Green Space” from the grid point.

Each grid point's urban land type is identified using the “Spatial Join” tool with the grid points as the target feature and the join feature as the urban typology data. The percentage of imperviousness of the land at each grid point is also similarly found using the “spatial join” tool. With the square grid as the area of influence of these grid points,

various other features for training the ML model, such as the built-up imperviousness, public or not public area, the number of trees, and the type of the nearest street, are also derived using geoprocessing tools like “Near”, “Spatial Join”,” intersect”. The feature engineering of the data is updated in the attribute table of the grid point feature, on which this geoprocessing analysis is done in ArcGIS. A buffer is created on the boundary of the extent of the study, and the reference grid points are removed in this boundary buffer area to avoid redundancies in model training. Figure 4 demonstrates the process used for feature engineering in ArcGIS. The percentage area of the square grid covered by the Crown projection area of the trees is calculated in ArcGIS using the “intersect” and “Dissolve” tools.

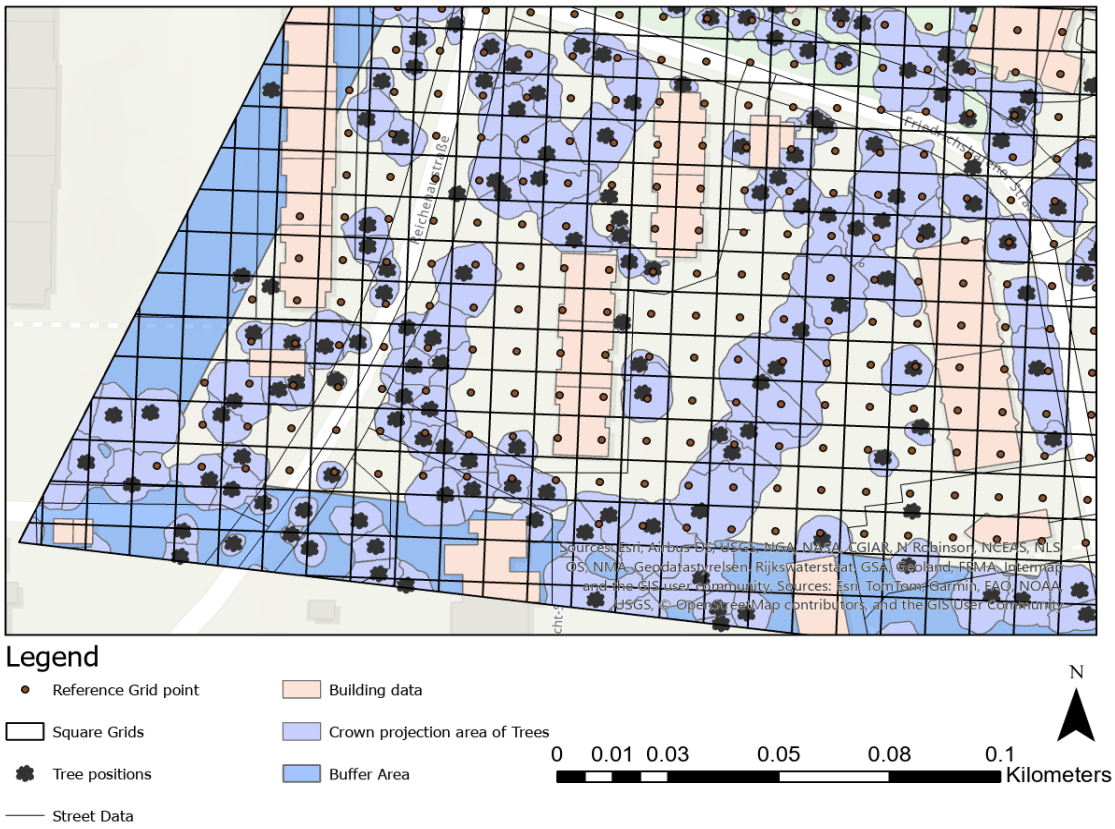

Figure 4 | GIS illustration of the processes used for feature engineering

3.3 Data Processing

This stage of the methodology involves making structured datasets compatible with training ML models by aligning the fields and correcting missing or error values from the datasets to avoid redundancies and deviations in the results. The attribute table of the reference grid point on which the feature engineering is done in ArcGIS is exported into

Excel format. Data processing is done in the Microsoft (MS) Excel sheet. Refer to Figure 5 for insight into the operation environment for various methodology stages. In data processing, all the deviating values and null values from the Excel sheet are removed. The derived feature-engineered data from the ArcGIS data contains many unwanted fields and null values. These noise data are removed and processed before it is used for model training.

The derived Excel data is processed in a series of steps. The first approach is to delete all the fields which are not needed for the study. The main fields required for the study are based on the derived feature-engineered data from ArcGIS.The Feature Engineered data from the ArcGIS is imported into the MS Excel sheet using the tool “Table to Excel”. Some new fields are added to the imported Excel sheet based on the feature engineering data. A field defining whether the grid point lies within a building is added to the derived data. The field specifying if the point is within the building is manually defined in the Excel sheet by comparing the values in the field “distance of the point to the nearest building”. All entries with a 0 value in the “distance of the point to the nearest building” signify that the point lies in the building. This way, the value is defined for the field “within a building or not”. Numerical data 1 or 0 is given for this data in the purview of model training. All the unwanted fields and features are removed in this process, and only the data required for model training is retained. The significant fields retained for the model training data from the list of fields from feature-engineered data are shown in Table 1 (Given below are the features which are found relevant based on the study):

Table 1 | Features used for model training

Feature Description	**Data Type**
Unique identifier for each reference point feature.	Numerical
X-coordinates of the point.	Numerical
Y-coordinate of the point.	Numerical
Distance from the point to the nearest street.	Numerical
The type of nearest street to the point,	Numerical/Categorical
Field specifying whether the point is within a building	Numerical
Distance from the point to the nearest building.	Numerical
Distance from the point to the nearest tree.	Numerical
Number of trees within the square grid.	Numerical
Height of trees within the grid.	Numerical
The scale imperviousness of the location of the point of the study	Numerical

The field specifies whether the location of the point of the study is public property.	Categorical
Classification of Land Use Type / Urban Typology	Numerical/ Categorical
The distance to the nearest cycle track	Numerical
The crown area of the nearest tree	Numerical
The angle of the location of the nearest tree concerning the reference point	Numerical
The distance to the nearest green or open space	Numerical
The built-up imperviousness feature of the reference point	Numerical
Greening potential of the grid (target variable for the model).	Numerical

3.4 Model Selection and Training

The decision to select and train the model is grounded in analysing the state of art section. Choosing the model for training is crucial since different models work for various data types and purposes. TensorFlow libraries are selected to train the machine learning model. The Structured data created with the feature engineering process in ArcGIS are suitable for ML tasks like classification and regression. To identify the complex patterns, relationships and the nature of large datasets used for training, the Feedforward Neural Network (FNN) model was used as the primary model training framework. The tensorFlow ML algorithm library builds and trains the model to learn the relationship between the city's spatial and physical features. Python script is used to input the ML models in the Integrated Development Environment (IDE), Google Colab.

The significant steps involved in building the model are

- Feature and Target Definition:

  The feature-engineered structured data from the Excel sheet in the IDE is loaded into the model. The features of the data are defined in the model. The training features are defined in terms of “x” and “y” variables. “x” is set as the feature matrix, and the “y” variable is set as the target variable; here, it is “greening potential”.

- Data processing

  The data that require normalisation to ensure values remain in a valid range are normalised (Angle data contains negative angle values, and it involves normalisation to avoid errors in the prediction of the model). The data is then standardised to ensure

numerical stability and to scale the values to have a mean of 0 and a standard deviation of 1 to enable model training. The standardised dataset is then split into training and testing datasets. Here, 80% of the dataset is used for training the data, and 20% is divided for testing the model for evaluation.

- Defining the Model

  The FNN is defined as having three hidden layers and one layer for regression. Dropouts are introduced in the model to regularise the model by preventing overfitting. Activation functions are also defined to enable non-linear learning of the neurons. Callback functions are also defined along with the model definition stage to avoid the overfitting of the model

- Model Compilation and Training

  The model is compiled with suitable optimizer and loss functions to improve the quality of the model's prediction and think about the measured predictions made. The model is compiled here with the “Adam’ optimiser and the mean_squared_error (MSE) loss function. After model compilation, the model is trained for different epoch combinations, batch sizes and validation splits.

- Model Evaluation

  The model’s performance is then evaluated based on the test dataset and the predictions made on the test dataset. The model’s performance is evaluated using the MSE and R^2 score.

Refer to Appendix 1 for the script used in this study.

3.5 Model Testing

The model is trained in the IDE, and the results of the model's predictions are evaluated based on the MSE and R^2 score. These values are assessed, and their prediction distribution is plotted using Python. The SHAP values are derived from fine-tuning and explaining the model to improve the prediction. The prediction data are then compared with the actual data to test the accuracy and improve the model’s performance. The accuracy and predictions of the model are cross-validated using GIS tools and data. Model improvement is done for different iterations till a suitable learning score, such as the R^2 score, is achieved.

3.6 Model validation

The model is validated using a Case study of the climate-neutral housing project at St. Michael Strasse (Maikäfersiedling), Munich. The data is prepared for the case study region with all the physical features except the target variable (greening potential), and this data sheet is fed to the model for prediction. The prediction results from the model are compared with the actual greening potential values of the validation study site. The prediction data is imported into ArcGIS to visualize and analyse the model's predictions to gain insights into the factors influencing tree planting locations in the city—the option for deployment of the model for practical usage.

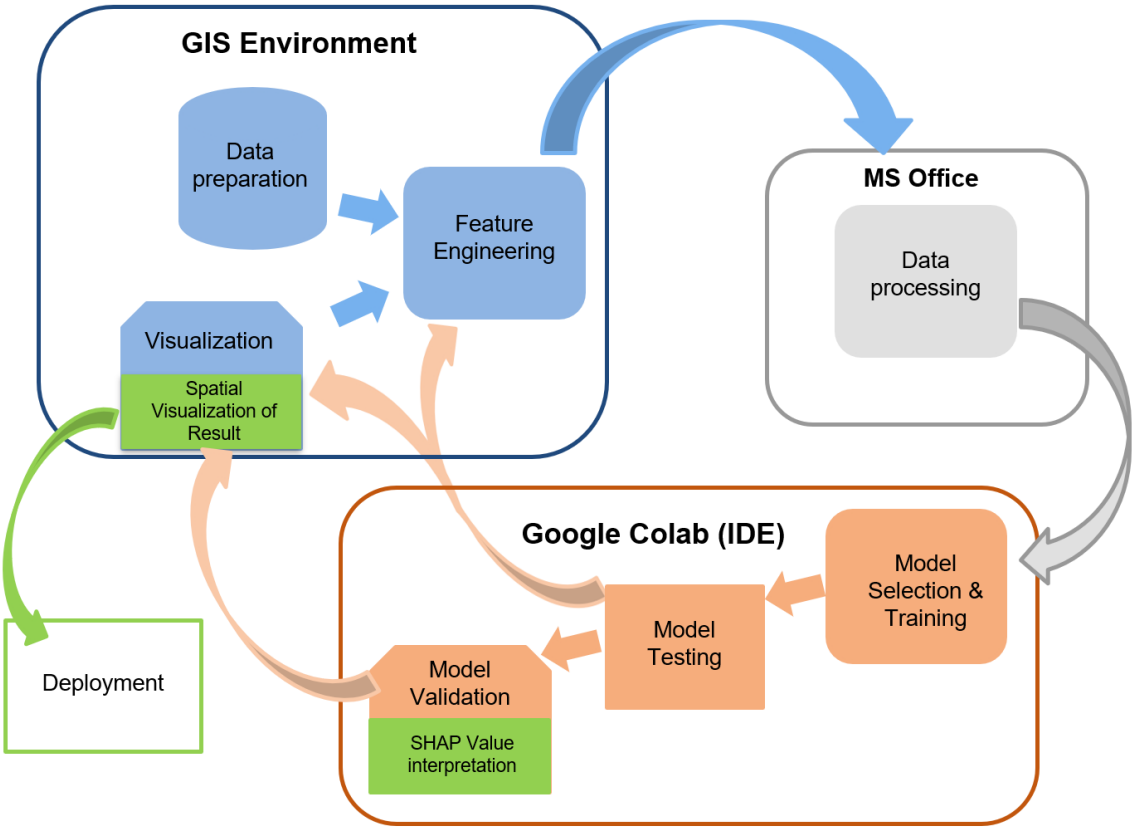

Figure 5 | Operation environment for various methodology tasks

3.7 Visualization & Deployment

The visualisation is done in the GIS platform. The predicted data from the model and the reference point coordinates are converted to the CSV format using Python scripting. This CSV format is imported into the ArcGIS map using the quick import tool in the same map in which the data set was prepared. The projected coordinate system is set to

ETRS 1989 UTM 32N, which can locate the points to the exact coordinates for which it is used. This trained model is then saved along with the architecture and weights of the model for future use.

4. Case Study

The case study of Munich in Germany is used to accomplish the study's objectives by applying the methodology steps. Munich has a good pool of GIS-compatible above-ground data about the various physical parameters of the city, which can be used to train ML models to learn the patterns and relationships between various physical features to answer the study's research questions. The study is divided into the pilot, the main, and the validation studies to achieve the desired level of accuracy of the study objectives. The ML model is trained based on the dataset derived as per the methodology, and different datasets are used in the three above-mentioned sections of the study. Figure 6 represents the flowchart of the progress of the study done under this case study.

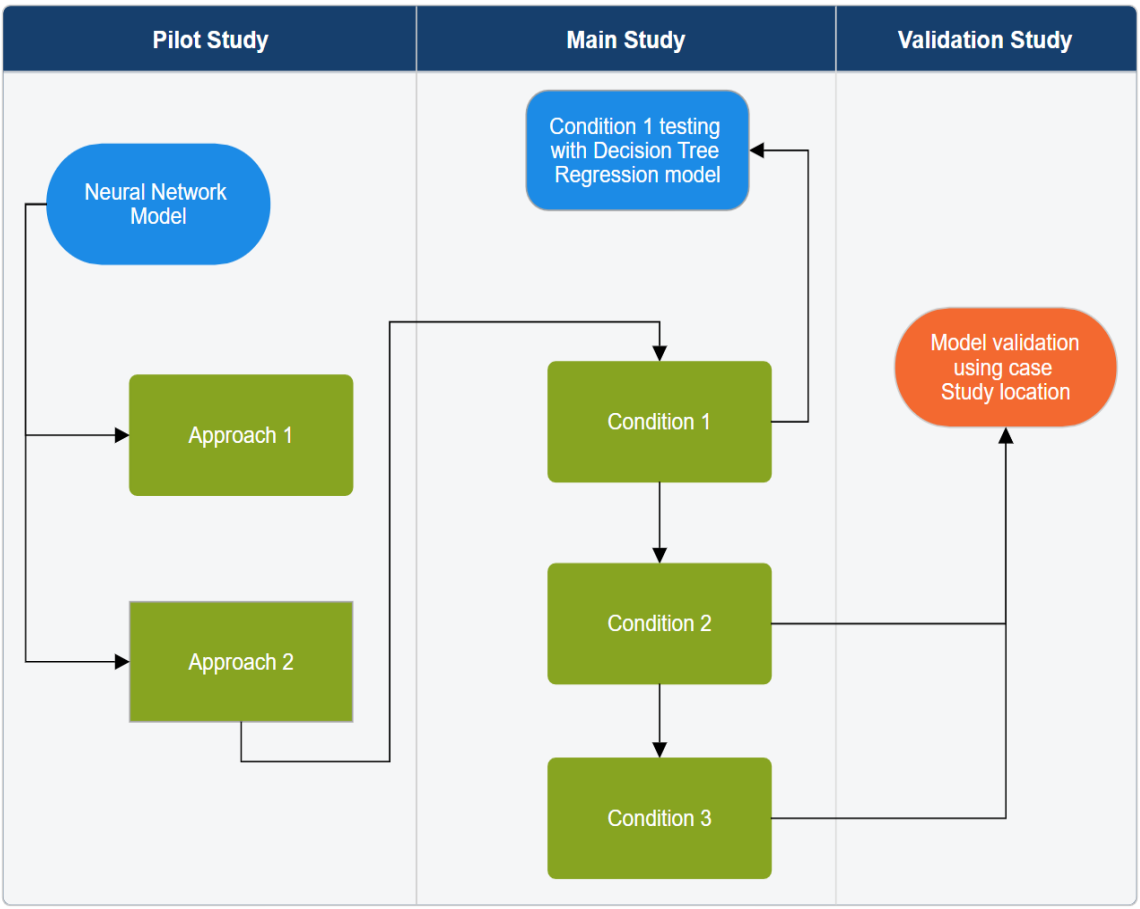

Figure 6 | Model training flow chart showing various studies done under the case study

4.1 Case Study Location & Data Used

Munich city has a good proportion of UGI and uniform-height buildings, and the availability of spatial, meteorological and physical data is chosen as the study area [97]. Munich is the third-largest city in Germany and the capital of the Bavarian state of Germany. Its population is over 1.4 million, and the area is around 311 km2. The site considered for the case study is located in Munich (48◦8' N, 11◦35' E, elevation of 520m above sea level). The mean annual temperature of Munich is 9.6 °C. Munich has both cold and warm months, with an average coldest temperature of 0.3°C in January, while the warmest temperature averages 18.9°C in July. The precipitation averages can reach a minimum of 46mm in January, and the winter in Munich is often drier. [98, 99]. Countries like Germany are preparing a system for effectively managing and developing their tree inventory using a tree information system. Munich in Germany has conducted multiple remote sensing and photogrammetry to create GIS data sets for effective urban management and development. Munich is developing 3D cityGML files with higher levels of Detail, and many of the GIS data are available in open source platforms. [22, 100–102]. Figure 7 shows the administrative boundaries of the case study area, Munich.

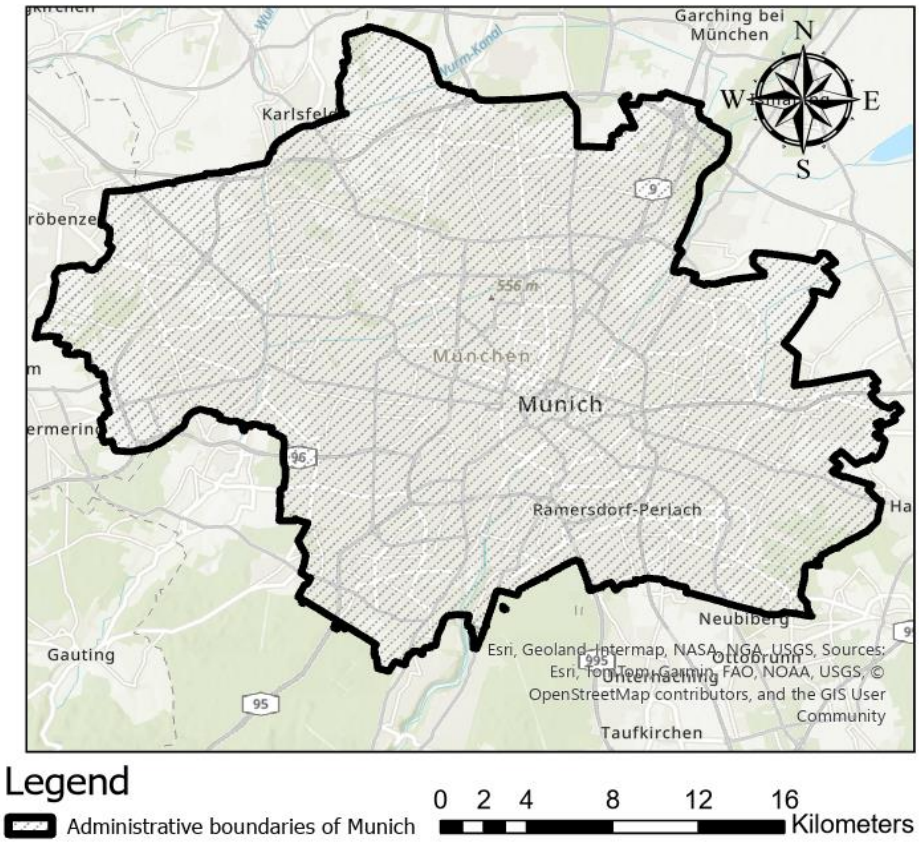

Figure 7 | Case study location, Munich, marked with administrative boundaries

Data used for the study

Further to the state-of-the-art section, various above-ground spatial and physical data of Munich city are gathered from different sources. The various physical data of Munich, such as the tree, surface imperviousness data, building data, built-up area imperviousness data, street data, cycle track data, data on public areas and land-use typology data, are the significant datasets used to derive the dataset for model training. The tree data used are the geodatabase files containing tree position and the tree segments showing the delineation of the crown projection area of the trees of Munich city region, which is created using remote sensing by Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) [100]. The surface imperviousness data, available in raster file formats with a resolution of 10m, is derived from the open-source platform Copernicus Land Monitoring Service [103]. The building data containing the vector data of Munich's buildings is derived from Munich's open-source data portal [104]. The cycle track data is also derived from the Open Data portal of Munich [105]. The street data is vector line data of the streets derived using Overpass Turbo, a web-based query tool for extracting specific data from the OpenStreetMap open-source database [106]. The vector data on the spatial information on the public area of Munich area is derived from the open data portal called “Geodatenservice der Landeshauptstadt München” [107]. The built-up area imperviousness data, also available in raster file formats with a resolution of 10m, is derived from the open-source platform Copernicus Land Monitoring Service [108]. The Urban typology of Munich containing information on the various classifications of land based on their use and type are availed from the research studies of the Research Training Group for integrated urban planning studies done under Technical University of Munich [101].

4.2 Pilot Study

A pilot study was conducted to train the model in defining the approaches for devising the approach for defining the target variables for prediction and analysing the outputs to affirm the methodology of the study. The pilot study aimed to get an overview of the framework and methods devised to make the AI ML Model. As the name suggests, a smaller area of 16.06 sq. km was used for the study, with only some prominent physical attributes considered for the feature engineering of the dataset to train the ML model.

The features considered for creating the structured feature engineering data for the model training consisted of Tree, Street, Urban Typology and Building data. With these data, the data was prepared and processed in the GIS tool using the various geoprocessing tools in ArcGIS, as explained in the methodology section. Two approaches were used to train the model in the pilot study in defining the target variables. The greening potential was set as the target variable, and two methods were used to describe it. The features of the structured feature engineering data are given below in Table 2, used in the model training and their data type. The features used in the model training of the pilot study are shown in the Table. Basic Sequential NN with different combinations of the number of Neuron layers with a dropout of 0.3 were used to optimise the results. Callback functions were not used in the pilot study, and 50 epochs were used in training the ML model. Figure 8 shows the area chosen for the pilot study, including the extent and the data used.

Table 2 | Features used in the Pilot study

Feature Description	**Data Type**
Unique identifier for each reference point feature.	Numerical
Unique identifier for the square grid enclosing the point feature	Numerical
X-coordinates of the point.	Numerical
Y-coordinate of the point.	Numerical
Latitude of the point.	Numerical
Longitude of the point.	Numerical
Unique identifier of the nearest street from the point.	Numerical
Distance from the point to the nearest street.	Numerical
Unique identifier for the nearest building from the point.	Numerical
Indicator of whether the point is within a building	Numerical
Distance from the point to the nearest building.	Numerical
Unique identifier for the nearest tree from the point.	Numerical
Distance from the point to the nearest tree.	Numerical
X-coordinate of the nearest tree from the point.	Numerical
Y-coordinate of the nearest tree from the point.	Numerical
Number of trees within the square grid.	Numerical
The total area covered by individual tree canopy	Numerical
Height of trees within the grid.	Numerical
Greening potential of the grid (target variable for the model).	Numerical
Classification of land use type (e.g., residential, commercial, etc.).	Categorical

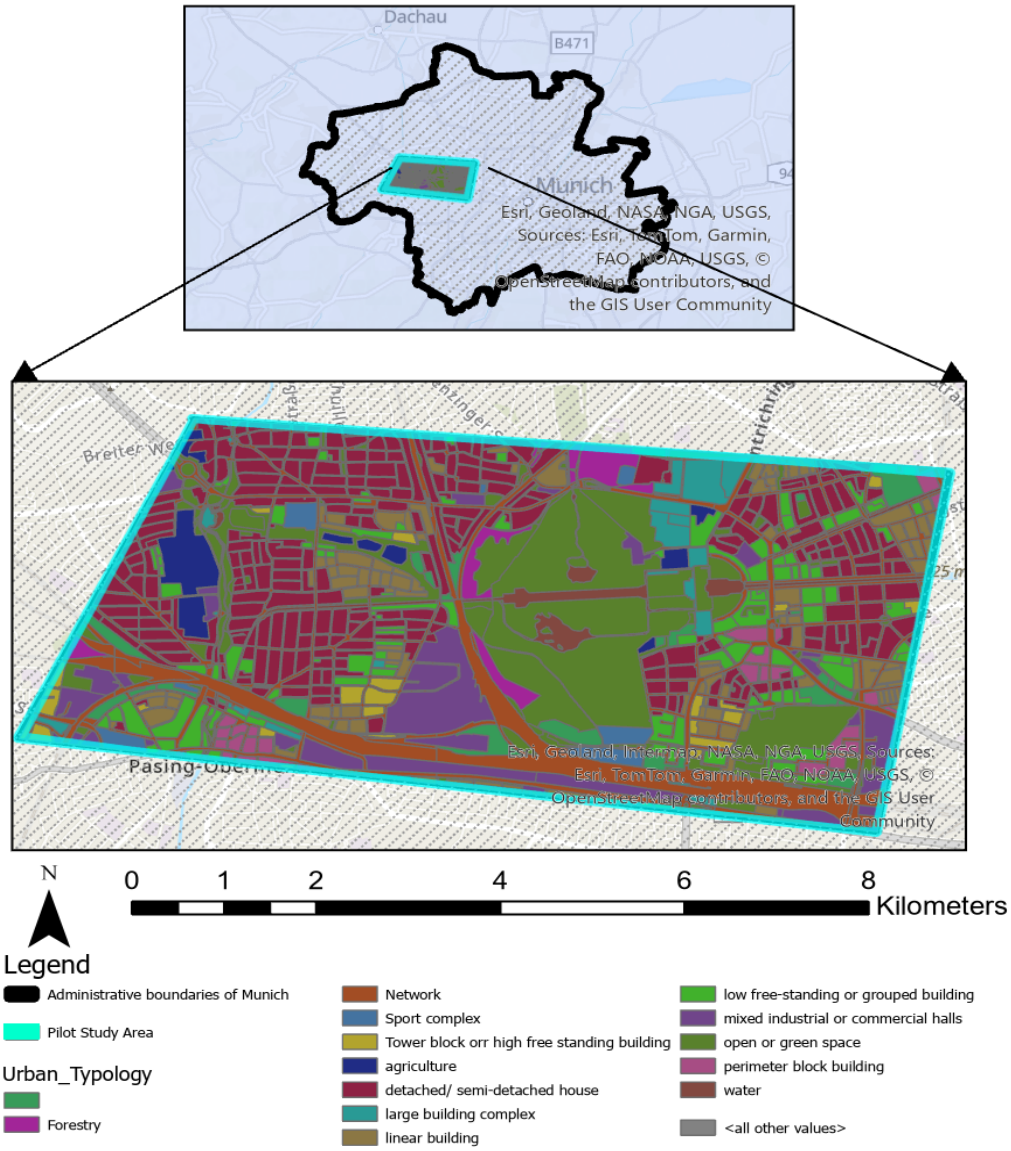

Figure 8 | Area selected for Pilot study and the data used in pilot study

4.2.1 Pilot Study Approach-1

In this approach for defining the target variable, an extra framework was to be performed manually on the dataset that was to be used for model training to define the values of the target feature “greening potential”. The framework created a set of conditions concerning the feature-engineered data. In this approach, some extra analysis in ArcGIS was also done to identify which land types (Urban typology feature) had a higher density of trees. A dimensionless number is derived for the density of trees on various kinds of land in Munich to identify the types of land having a significant share of trees compared to others. It was derived using a set of operations in ArcGIS with the “spatial join” tool and adding a new field with mathematical operations in the attribute table of

the “Urban Typology” shapefile. This dimensionless number is calculated by the number of trees enclosed in each type of land divided by the area of each land type chosen for the “spatial join” analysis. The derived values for various land types are given in Table 3. Based on this dimensionless Tree density number and feature engineering data, a set of conditions was introduced to define the target variable for model training “greening potential”. According to the first approach, three values were determined for “greening potential”, which are 0.9, 0.5 and 0.1. The conditions used to define these values can be found in Table 4. Refer to Appendix 2, figure 26 for the dataset used for model training for this approach.

Table 3 | Classification of various land typologies present in Munich with their corresponding dimensionless tree density number

Various land types under the feature Urban typology	Dimensionless Tree density number
mixed industrial or commercial halls	2.59
Network	3.32
low free-standing or grouped building	4.9
perimeter block building	3.81
open or green space	5.58
large building complex	3.92
water	3.81
Tower block or high free-standing building	4.8
linear building	5.41
Sport complex	2.83
None	5.82
detached/ semi-detached house	6.9
Forestry	8.84
agriculture	0.88

Table 4 | Categories of Greening potential values used in Pilot study Approach 1

Condition 1: High Greening Potential = 0.9	The point is not within a building.
	There are no buildings near the point in the enclosed square grid.
	There are no streets near the point in the enclosed grid.
	The number of trees within the enclosed grid is greater than or equal to 1
	Land usage type is forestry, Open or green space, detached/ semi-detached house
Condition 2: Medium Greening Potential = 0.5	The point is not within a building.
	Distance to the nearest building from the point is greater than 3m.
	Distance to the nearest street from the point is greater than 2m.
	The number of trees within the enclosed grid is greater than or equal to 0
	Land usage type is all except agriculture.
Condition 3: Low Greening Potential = 0.1	Rest of the condition

4.2.2 Pilot Study Approach 2

In the first approach, the analytical framework for defining the greening potential required careful analysis of the feature engineering data. The flexibility of feature engineering data is also challenging, as it can differ for different cities. The approach

also needs a lot of pre-analysis to be done by humans on the feature engineering data, which undermines the potential of the ML AI models.

In the second approach, the “greening potential” target variable was defined using the tree segments data, which defined the crown delineation of the trees. The greening potential for each point is determined by the percentage of the square grid area of the respective reference grid point covered by the crown projection area of the trees enclosed in the respective grid. Figure 9 gives a visual illustration of how this approach is done in the GIS platform.

$$Greening\ potential = \frac{Crown\ projection\ area\ of\ the\ trees\ in\ a\ grid}{Area\ of\ the\ Square\ grid} \times 100$$

For each grid, the greening potential value is derived using geoprocessing tools such as the “intersect” and “dissolve” tools in ArcGIS. This way, the “greening potential” is derived for every point, and this data is used as a target variable for training the model. This approach requires less human intervention in defining the greening potential and can fit into the framework for further studies in different study locations.

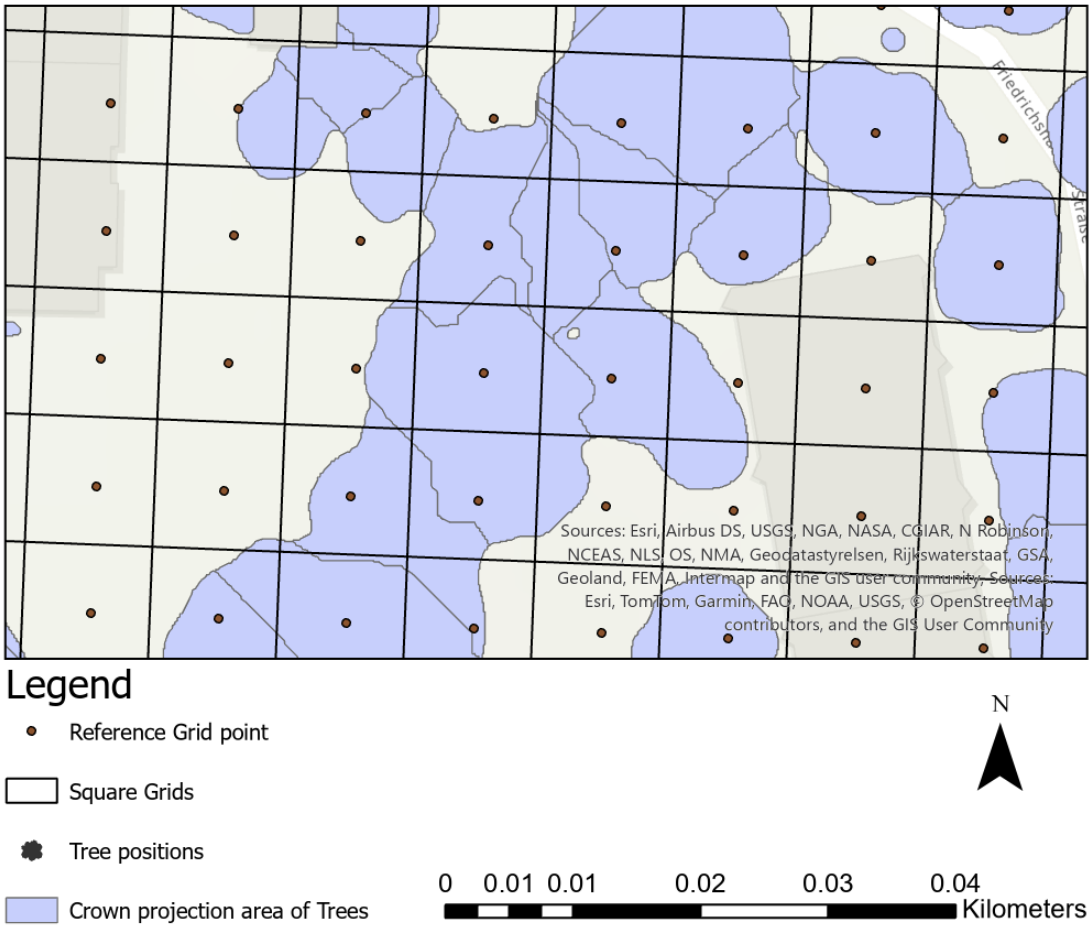

Figure 9 | Illustration of Greening potential definition used in Pilot Study Approach 2

This approach of defining the Greening potential is used in the coming sections of the study to train the model for further model optimization and achieving the objective. This approach involves more dependence on GIS and ML and better integrates the purpose of the study framework.

4.3 Main Study

The main study is the major area of focus, for training the model and optimizing it. The motive of the main study is to establish the framework suggested in the study's objective with concrete findings and build a model capable of learning the physical and spatial relationships of various city attributes. A comparatively larger area of 47.8 sq.km was used to derive the dataset, encompassing all the relevant GIS-based data available for Munich. The data for creating the structured feature engineering data for the model training consisted of tree data, building data, street data, urban typology data, cycle track data, surface imperviousness data, built-up imperviousness data, and vector data defining public properties. With these data, the data was prepared, feature-engineered and processed using the various geoprocessing tools in ArcGIS, as explained in the methodology section. The study area is depicted in Figure 10, showing the extent and data used. The target variable (“greening potential”) is determined by the percentage of the square grid area of the respective reference grid point covered by the crown projection area of the trees enclosed in the respective grid as mentioned in the second approach of the Pilot Study. Refer to Appendix 2, Figure 25, for the dataset used for model training.

Different iterations of the model training are done with different features and with different approaches to avoid overfitting problems and improve the performance of the model. Of the various iterative models used, 3 different conditions of model training approaches, which resulted in the better performance of the model comparatively, are discussed in this section. The SHAP values are derived for each condition of the model training. Each condition is ranked in ascending order in terms of improvement in architecture and performance of the ML model, i.e., the condition 3 model is an improved version of the condition 1 model. Methods to reduce overfitting were introduced in this study, such as early stopping functions, regularization, optimized feature engineering based on SHAP interpretations

The main training features retained from the feature-engineered dataset after multiple iterations of model optimization are given in Table 5.

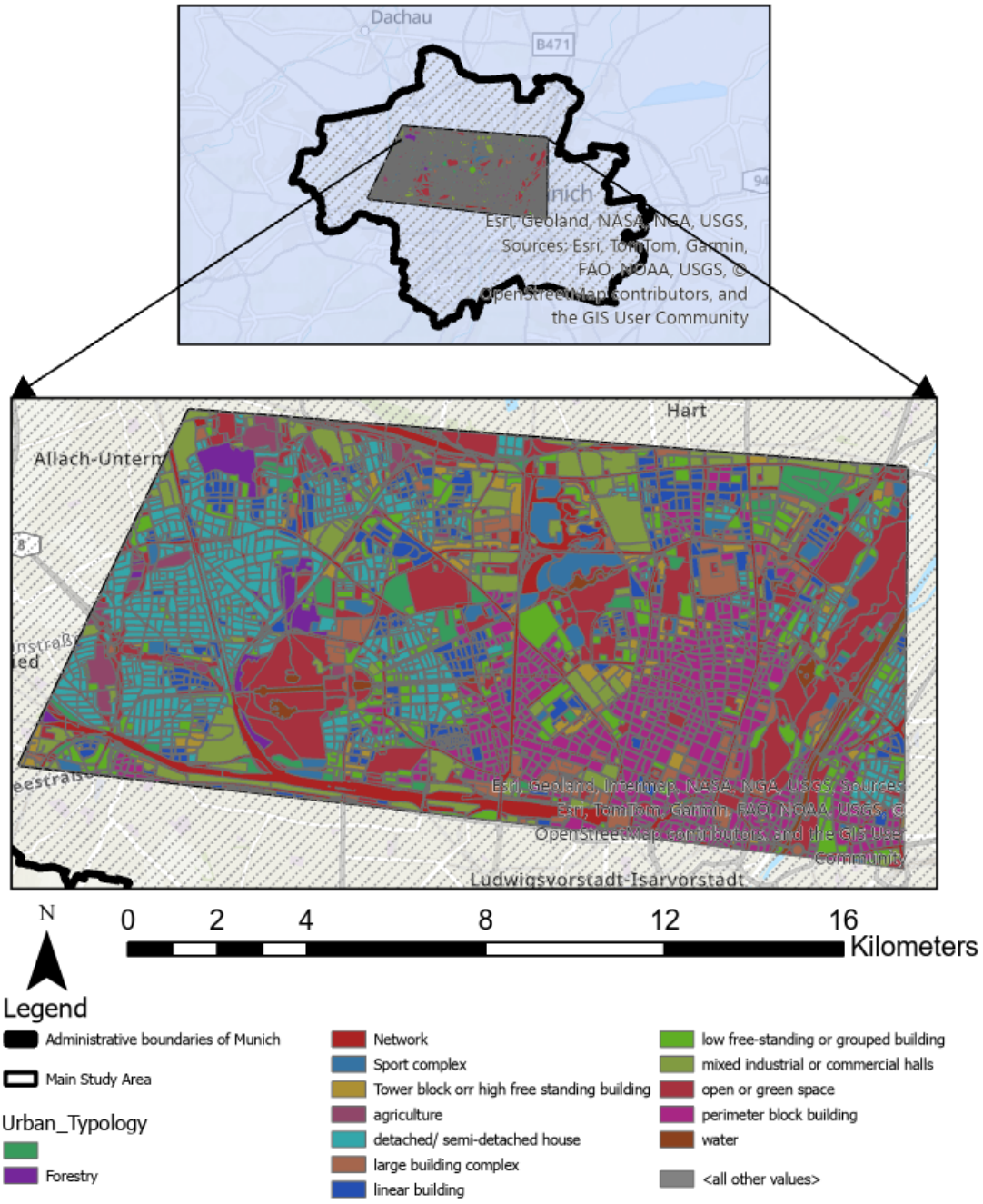

Figure 10 | Area selected for the Main study and the data used in the Main Study

Table 5 | Features used in the main study

Feature Description	Data Type
Unique identifier for each point feature.	Numerical
X-coordinates of the point.	Numerical
Y-coordinate of the point.	Numerical
Number of trees within the square grid.	Numerical
Distance from the point to the nearest tree.	Numerical
The crown area of the nearest tree	Numerical
Height of trees within the grid.	Numerical
The angle of the location of the nearest tree with respect to the reference point	Numerical
Field specifying whether the point is within a building	Numerical
Distance from the point to the nearest building.	Numerical
The built-up imperviousness feature of the reference point	Numerical
Distance from the point to the nearest street.	Numerical
The type of nearest street to the point	Numerical/Categorical
The distance to the nearest cycle track	Numerical
The scale imperviousness of the location of the point of the study	Numerical
Classification of Land Use Type / Urban Typology	Numerical/ Categorical
The distance to the nearest green or open space	Numerical
Field specifying whether the location of the point of the study is public property	Categorical
Greening potential of the grid (target variable for the model).	Numerical

4.3.1 Condition 1

In this feature-engineered dataset, there are 3 categorical data features considered: “the type of nearest street to the point”, “Classification of Land Use Type / Urban Typology”, and “field specifying whether the location of the point of the study is public property”. The rest of the data are numerical datasets. Machine learning data requires numeric data for computation and one-hot encoding is done on these categorical data to convert these data into formats that can be fed into the ML algorithms to learn the data to make predictions. These data sets are trained for the model, and the SHAP values are derived to analyse the features influencing the model predictions and performance. The predicted data is visualized and analysed in the GIS platform for detailed interpretation.

4.3.2 Condition 2

This condition is defined in the purview that the model can be tested on regions that may not encompass all land typologies or street types used for the training. The Features “the type of nearest street to the point” and “Classification of Land Use Type / Urban Typology” contain multiple distinct entry types. Refer to Table 3 under the Pilot Study section for insight on different land types under the feature Urban typology. Similarly, there are different types of streets defined in the Street data. Refer to Table 13 in Appendix 3 for insights on different street types described in the study data. The one-hot encoded categorical features compiled in the model training algorithm identify each subsection of the categorical feature (e.g. Urban typology) as separate features such as “open of green space”, “forestry”, etc. So, the categorical features containing multiple values of distinct strings are replaced with numerical data. NN can introduce unintended ordinal relationships among various values of features used in training due to the numerical encoding of the categorical features. Therefore, a judicious approach of numerical encoding of the categorical features “the type of nearest street to the point” and “Classification of Land Use Type / Urban Typology” are necessary to avoid this unintended generalization of the model. The Numerical encoding of the “Classification of Land Use Type / Urban Typology” is done according to the dimensionless tree density number in Table 3 in the pilot study section. For the Street type numerical encoding, first, the “street data” in ArcGIS is intersected with the “surface imperviousness”. Then, the street types are indexed with a maximum count of “surface imperviousness” values. Thus, the street type is mapped with surface imperviousness data. The determined numerical encoding of the street type is given in Table 13 in Appendix 3. These data sets are trained for the model and the SHAP values are derived to analyse the model features and performance. The predicted data is visualized and analysed in the GIS platform for detailed interpretation. This condition established the flexibility of the model to be applicable in different urban areas of the city.

4.3.3 Condition 3

In this stage of optimizing the model performance, the features that can make the model overfit the data are avoided to check the model's performance based on an iterative process. The features are avoided based on logical reasoning as well. The model training results and predictions from condition 1 of the main study and condition 2 of the main study area are analysed to find the features which led the model to overfit the data. After analysis, certain land types are avoided from the study area to train the model.

The avoided land types in the study area are “open or Green Space” and “forestry”. The data from these Land types in the study area are not used for model training to avoid overfitting the model. The “open or Green Space”, and “forestry” land types contained trees planted in certain patterns in certain regions, and these similar patterns are not reproduced or mirrored in the similar types of lands in other parts of the study areas. Another reason to avoid these land types is because these sites can be easily accessed for planting new trees for the government or the public. These data sets are trained for the model, and the SHAP values are derived to analyse the model's features and performance. The predicted data is visualized and analysed in the GIS platform for detailed interpretation.

4.4 Model Validation in Maikäfersiedlung

The case study location of the climate-neutral housing project at St Michael Strasse (Maikäfersiedling) was selected, and the extent of the project area of the case study was 0.35 sq. km. This project area is proposed to focus on climate-neutral neighbourhood studies, and one objective of the study is to identify suitable locations for tree planting in the neighbourhood. [109]. Hence, this study validated the model in this area on the prospect of insights for future studies with data from authorities.

The feature engineering dataset was prepared using the methodology section for this area. The Structured feature-engineered dataset of the validation case study area is prepared with all the features mentioned in the main study section, except the target feature, “greening potential”. The actual target variable, “greening potential,” values of the location are also derived in the data preparation process so that they can be used to compare with the predictions made. This dataset is loaded into the trained model to predict the “greening potential” at each point based on the input dataset. The output data is checked for model prediction performance, and the predicted data is visualized in the GIS platform to interpret the data. Figure 11 highlights the location, extent of the validation study area, and the data used to build the dataset for testing.

4.5 Comparison of the Model training using NN and Decision Tree Regression Model

The model training process used in the main study condition 1 is applied to another ML algorithm, the Decision tree regression model, to compare it with the performance of the NN model. This study was done to affirm the choice of the model used for this study. using the Decision Tree Regression model. Both the models are run with the same dataset prepared for condition 1 of the main study. The model performance is evaluated to identify the better-performing model.

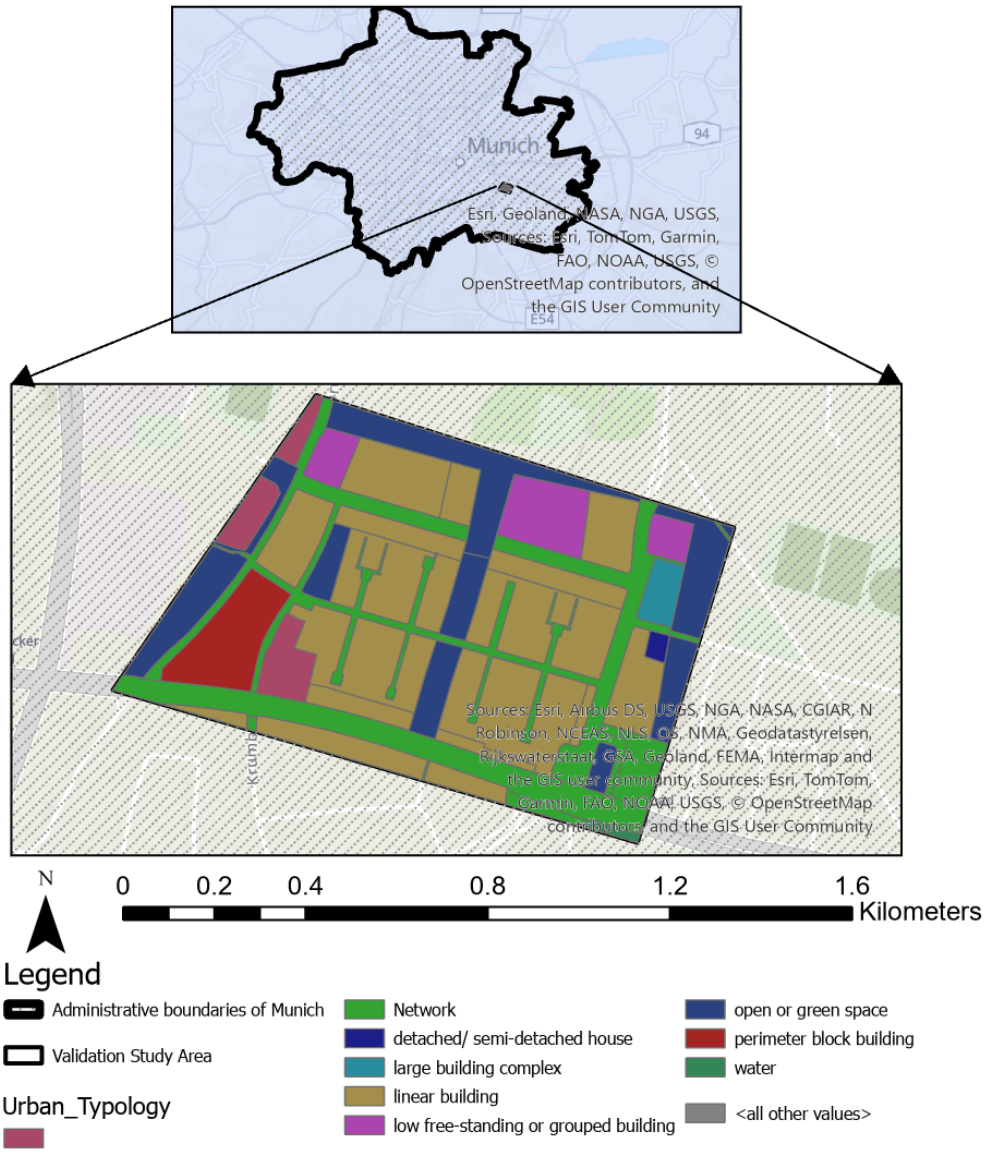

Figure 11 | Area selected for Validation study and the data used in Validation study

5. Results

After executing the model training, the outputs from the model training are interpreted using different plotting methods and results are documented. Different iterations of various studies and optimization of feature engineering based on the SHAP values and visualization are done to improve the performance. The results of various sections are discussed in different sections. Results are discussed in the order of the methodology and progress of workflow of the various study areas.

5.1 Results from the Pilot Study

The following section presents the outputs of the FNN model training used in the Pilot study, focusing on the model's performance and the evaluation of the outputs. The results of approach 1 of the pilot study are first discussed, followed by approach 2.

5.1.1 Results from Pilot Study, Approach 1

Model Training performance

The NN model was trained for 50 epochs and was configured to monitor the validation loss. The training performance shows that the "final training loss" achieved was 0.0037, and the "validation loss" was 0.0025. The lower values of "validation loss" and "training loss", along with the convergence of these values across the epoch of training, indicate that the model can effectively generalize the learning data without overfitting. Table 6 shows the training results from the pilot study approach 1.

Model Testing Evaluation

The model was evaluated on the test set, and the R^2 score given for the testing is 0.9632. This indicates that the model was able to explain approximately 96.32% of the variance in the target variable and a higher degree of prediction accuracy (greening potential). The "test loss" MSE value was found to be 0.0026, which is similar to the validation loss, implying the efficiency of model architecture in predicting the target variables more accurately.

Table 6 | Model Training Results for Pilot Study Approach 1

	Training Loss	Validation Loss	Test Loss (MSE)	R^2 Score
Output value	0.0037	0.0025	0.0026	0.9632

Interpretation of the Test Results with Plot Diagram

For a detailed interpretation and to interpret the model's performance from the pilot study, the prediction from the test data set is plotted against the actual values of the target variable (greening potential). A jitter scatter plot was plotted with the actual target variable values of the test dataset (Actual Greening potential) in the x-axis against the predicted values of the model (Predicted Greening potential) based on the test dataset. A 45-degree line was also plotted to interpret the uniformity of the prediction and actual greening potential values. Refer to the jitter scatter plot in Figure 12. In the scatter plot, since the actual greening potential values were restricted only to 0.9, 0.5,0.1, there was no continuous data distribution across the x-axis. The model predicted values outside the actual values of 0.9,0.5, and 0.1, with a continuous distribution of predicted values ranging from 0 to 1. In the scatter plot, a higher density of points was observed along the 45-degree line, implying better prediction values in line with the actual values. The distribution of the points away from the 45-degree line was of lesser density, suggesting that only some predictions have more significant deviations from the actual values.

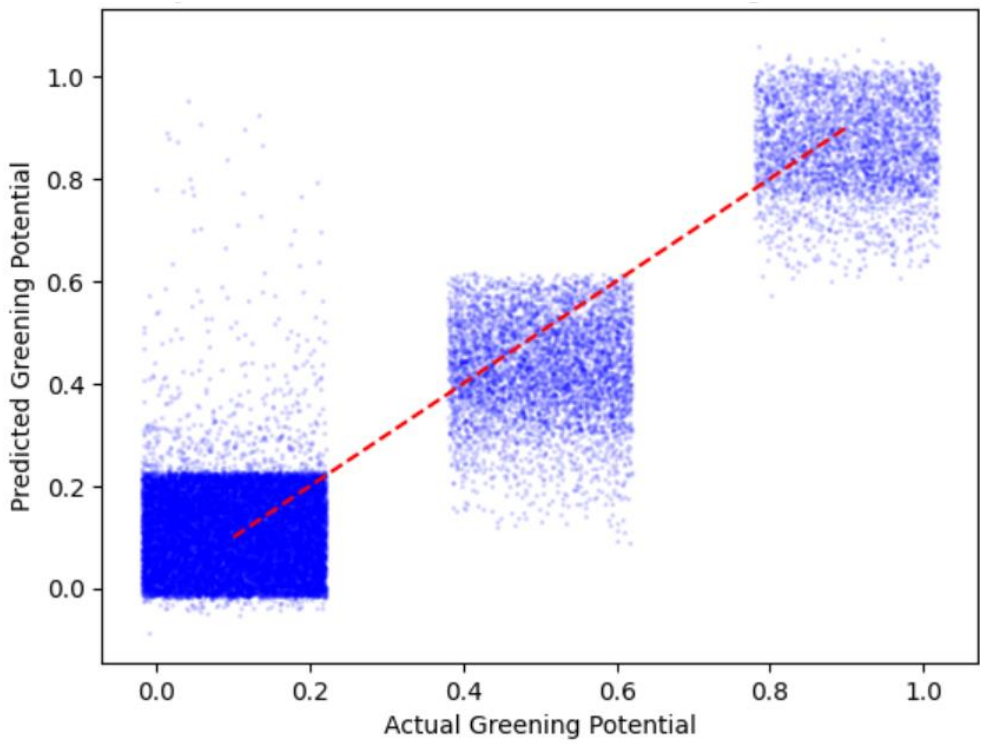

Figure 12 | Jitter Scatter Plot: Actual vs. Predicted Greening Potential values for pilot study approach 1

5.1.2 Results from Pilot Study Approach 2

Model Training performance

The methodology of defining the target variable (greening potential) defined in this study is the basis for all the studies to follow. Hence, the results of the studies are looked into to get maximum insight for improvement and optimization. This NN model was also trained for 50 epochs and was configured to monitor the validation loss. The training performance shows that the “final training loss” achieved was 618.47, and the “validation loss” was 609.46. The higher values of “validation loss” and “training loss”, along with less convergence of these values across the epoch, indicate the model is overfitting the learning data, and there is room for improving feature engineering data and model to avoid unwanted overfitting of the model. Table 7 shows the training results from the pilot study approach 2.

Model Testing Evaluation

The model was evaluated on the test set, and the R^2 score given for the testing is 0.6463. This indicates that the model was able to explain approximately 64.63% of the variance

in the target variable and suggests moderate performance, indicating partial alignment of the prediction with actual values (greening potential). These results suggest improvement in the overall model building and dataset preparation. The “test loss” MSE value was 591.62, and slight convergence with validation loss suggests the average performance of the model's architecture. These results indicated the need for further optimization.

Table 7 | Model Training Results for Pilot Study Approach 2

	Training Loss	**Validation Loss**	**Test Loss (MSE)**	**R^2 Score**
Output value	618.47	609.46	591.62	0.6463

Interpretation of the Test Results with Plot Diagram

The scatter plot considers a continuous distribution of actual greening potential values ranging from 0 to 100 for model training. The model predicted values ranging from 0 to 100 as well. A scatter plot was plotted with the actual target variable values of the test dataset (Actual Greening potential) in the x-axis against the predicted values of the model (Predicted Greening potential) based on the test dataset. Figure 13 shows the scatter plot diagram showing the distribution of the results in the graph. The scatter plot had a comparable concentration of point distribution along the 45-degree line. However, many outliers suggested more refinement of the training dataset and optimizing the model. Also, there was a high degree of over-prediction and underprediction, as observed from the graph, where there is a higher concentration of points lying as outliers at actual potential values of 0 and 100.

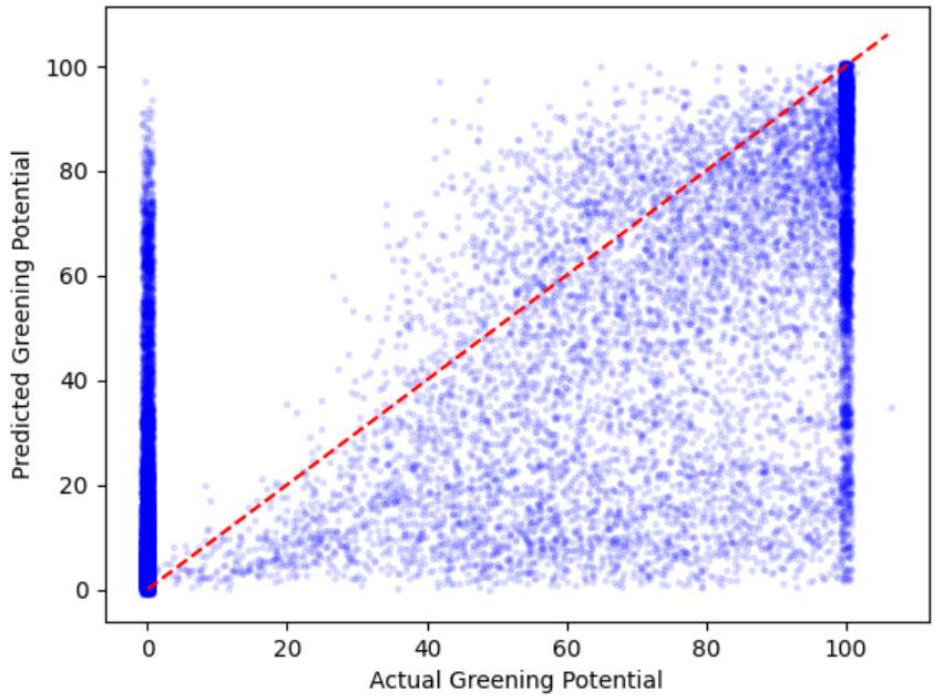

Figure 13 | Scatter Plot: Actual vs. Predicted Greening Potential values for pilot study approach 2

5.2 Results from the Main Study

The following section presents the outputs of the FNN model training used in the main study, focusing on the model's performance and the evaluation of the outputs. The results of various conditions of the main study are discussed in the order of condition 1, condition 2 and condition 3, respectively.

5.2.1 Results from Condition 1 model training

Model Training performance

This NN model was trained with call-back functions to avoid overfitting and was configured to monitor the validation loss. Advanced optimization techniques like regularization, early stopping function, and one-hot encoding were utilized to prevent overfitting techniques such as regularization. The training performance shows that the "final training loss" achieved was 312.06, and the "validation loss" was 312.69. The "validation loss" and "training loss" converge steadily. Similar training and validation loss indicate the model has better architecture, and the model was capable of learning data

without much overfitting compared to the model trained in the pilot study approach 2. Table 8 shows the model training results for the main study, condition 1

Model Testing Evaluation

The model was evaluated on the test set, and the R^2 score given for the testing was 0.7716. This indicates that the model could explain approximately 77.16% of the variance in the target variable and had good performance, indicating better alignment of the prediction data with actual target values (greening potential). This result suggests further scope for improvement in the overall model building and dataset preparation. The “test loss” MSE value was 311.52, and convergence with validation loss suggests better performance of the model's architecture without much overfitting. The presence of the multiple categorical features in this model and their one-hot encoding approach to make the data compatible with the model was found to add up the input shape size of the NN model, which envisaged the lesser flexibility of the model in the application of areas which do not contain all the categorical features like the study area in which the model was trained.

Table 8 | Model Training Results for Main Study, condition 1

	Training Loss	**Validation Loss**	**Test Loss (MSE)**	**R^2 Score**
Output value	312.06	312.69	311.52	0.7716

Interpretation of the Test Results with Plot Diagram

The SHAP values are derived for the model training for data exploration, debugging and explanation. A bee-swarm plot was derived, which gives an overview of all predictions with all the SHAP value features. The waterfall plot of a particular model prediction was also plotted to identify the feature dependencies in model training.

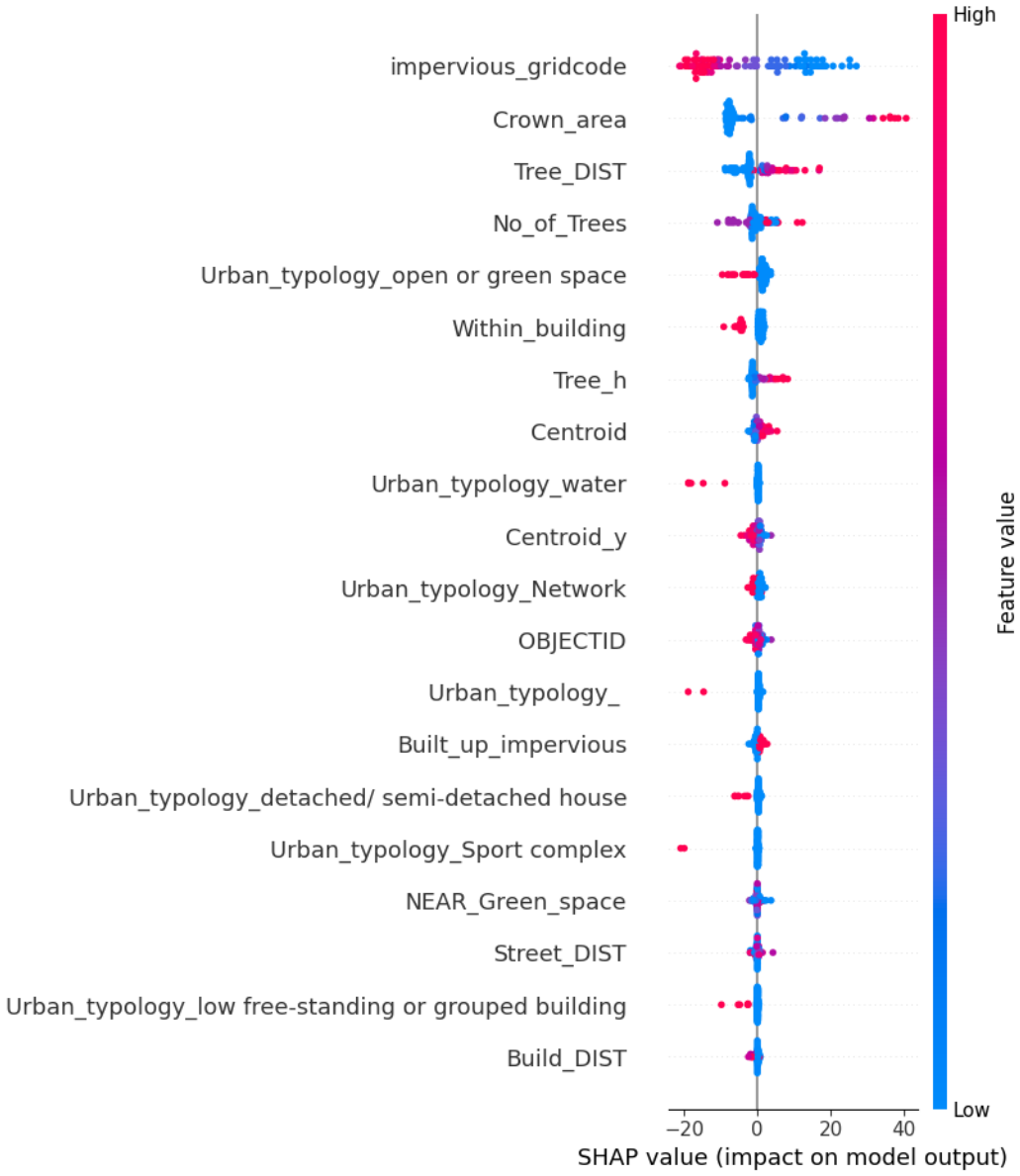

Figure 14 | Bee-swarm plot of SHAP values of all features showing an overview of prediction derived for model trained for condition 1 dataset of the main study

Figure 14 shows the bee swarm plot of the SHAP values for the condition 1 main study, indicating that the subsections of the categorical features are identified as separate features by the model due to one-hot encoding of the categorical features. In this case, the Urban typology features are now divided into multiple features like “Urban_typology_open or green space”, “Urban_typology_water”, “Urban_typology_network”, etc. This indicates the model is less flexible in other areas. The Features like surface imperviousness (impervious_ grid code), vegetation data like crown area (Crown_area), distance to the nearest tree (Tree_Dist), and building footprint (Within_building) are found to be the most significant factors influencing the

predictions with the consistent distribution of SHAP values. The higher surface imperviousness values negatively impacted the greening potential, while the higher values of the crown area positively impacted the prediction.

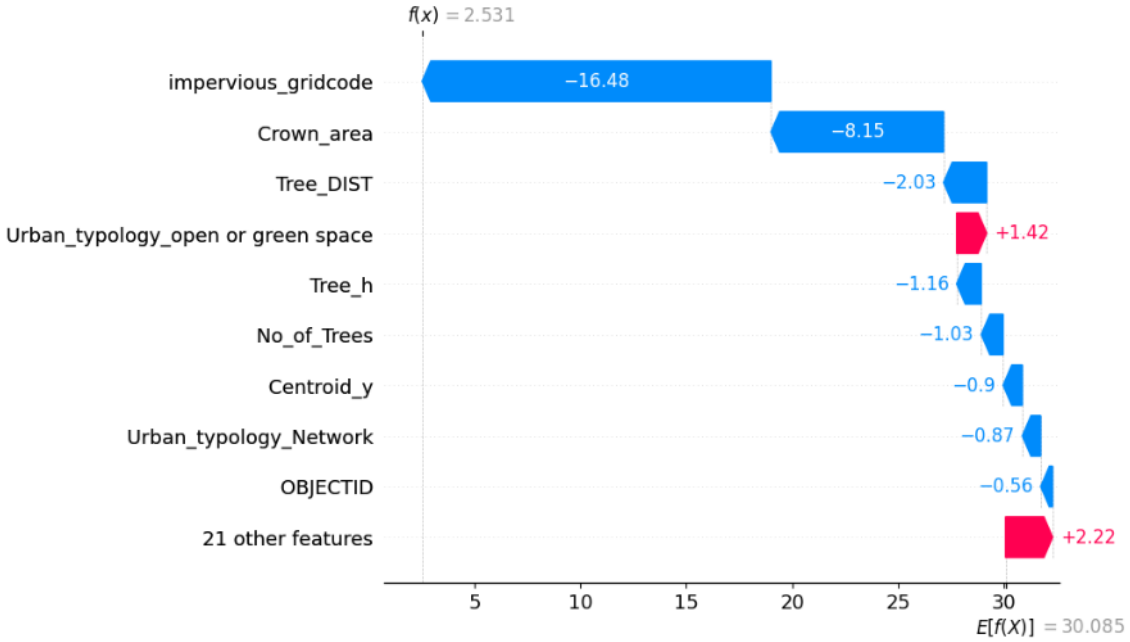

Figure 15 | Waterfall plot of SHAP values of a single sample of prediction derived for model trained for condition 1 dataset of the main study

Fig 15 illustrates the waterfall plot of the SHAP value of a single prediction, which exemplifies how the SHAP values of each contributing feature affect the prediction of the greening potential at that particular point. The analysis of the SHAP plot and the dataset of that particular point revealed that the point had higher imperviousness values and a low value of the crown area, along with other features, contributing to a lower greening potential for that point.

The predicted test data of the model is extracted into CSV file format and exported into the GIS platform for further visualization and analysis. The subtraction difference between the predicted greening potential and actual greening potential values of each point is visualized in ArcGIS for interpretation (Refer to Figure 16).

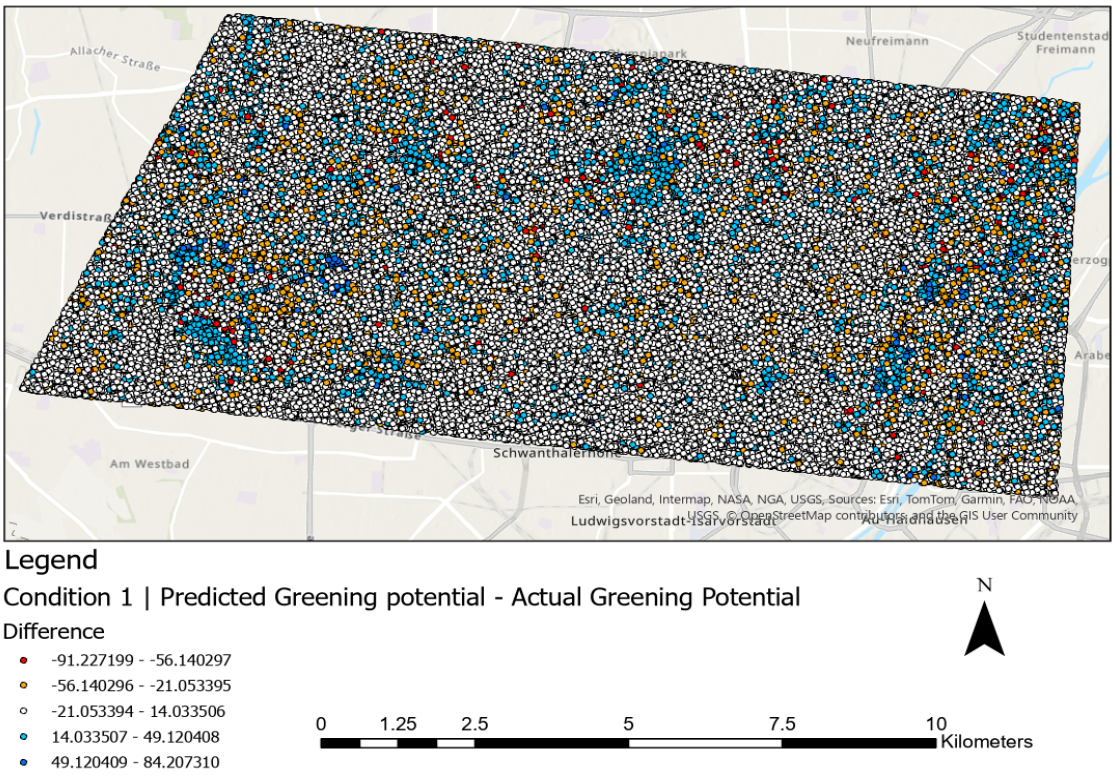

Figure 16 | Subtraction difference between the predicted Greening potential and Actual greening potential values of each point is visualized in ArcGIS for the model trained for Condition 1 of the main study

The test data prediction visualization helps to identify locations with over-predictions and under-predictions. The overprediction and underprediction areas were mainly found near the “open or green spaces” and “forestry” land typologies, which can be looked upon for further optimization efforts by thoroughly investigating the datasets with extra information and data.

5.2.2 Results from Condition 2 model training

Model Training performance

The NN model trained in condition 1 is not flexible enough to be tested for predictions involving small-scale areas, which do not encompass all the land typology and street types it was trained for. This condition is employed to increase the model's flexibility, in which the categorical feature data is converted to numerical data. To avoid overfitting, the NN model is trained with early stopping functions and regularization techniques such as dropout. It was also configured to monitor the validation loss. The training

performance shows that the “final training loss” achieved was 316.06, and the “validation loss” was 332.13. The “validation loss” and “training loss” value convergence was comparable to condition 1. Still, it was slightly less, suggesting the NN model has learned unintended relationships from the numerically encoded categorical data. Table 9 shows the model training results for the condition 2 of the main study.

Model Testing Evaluation

The model was evaluated on the test set, and the R^2 score given for the testing was 0.7557. This indicates that the model was able to explain approximately 75.57% of the variance in the target variable and good performance, indicating better alignment of the prediction data with actual values (greening potential). Still, its test performance is less than the condition 1 model. This result suggests improvement in the overall model building and dataset preparation. The “test loss” MSE value was 333.23, and convergence with validation loss suggests better performance of the model due to the robust architecture of the model.

Table 9 | Model Training Results for Main Study, condition 2

	Training Loss	**Validation Loss**	**Test Loss (MSE)**	**R^2 Score**
Output value	316.06	332.13	333.23	0.7557

Interpretation of the Test Results with Plot Diagram

The SHAP values are derived for the model training for data exploration, debugging and explanation. A bee-swarm plot was derived, which gives an overview of all predictions with all the SHAP value features. The waterfall plot of a particular model prediction was also plotted to identify the feature dependencies in model training.

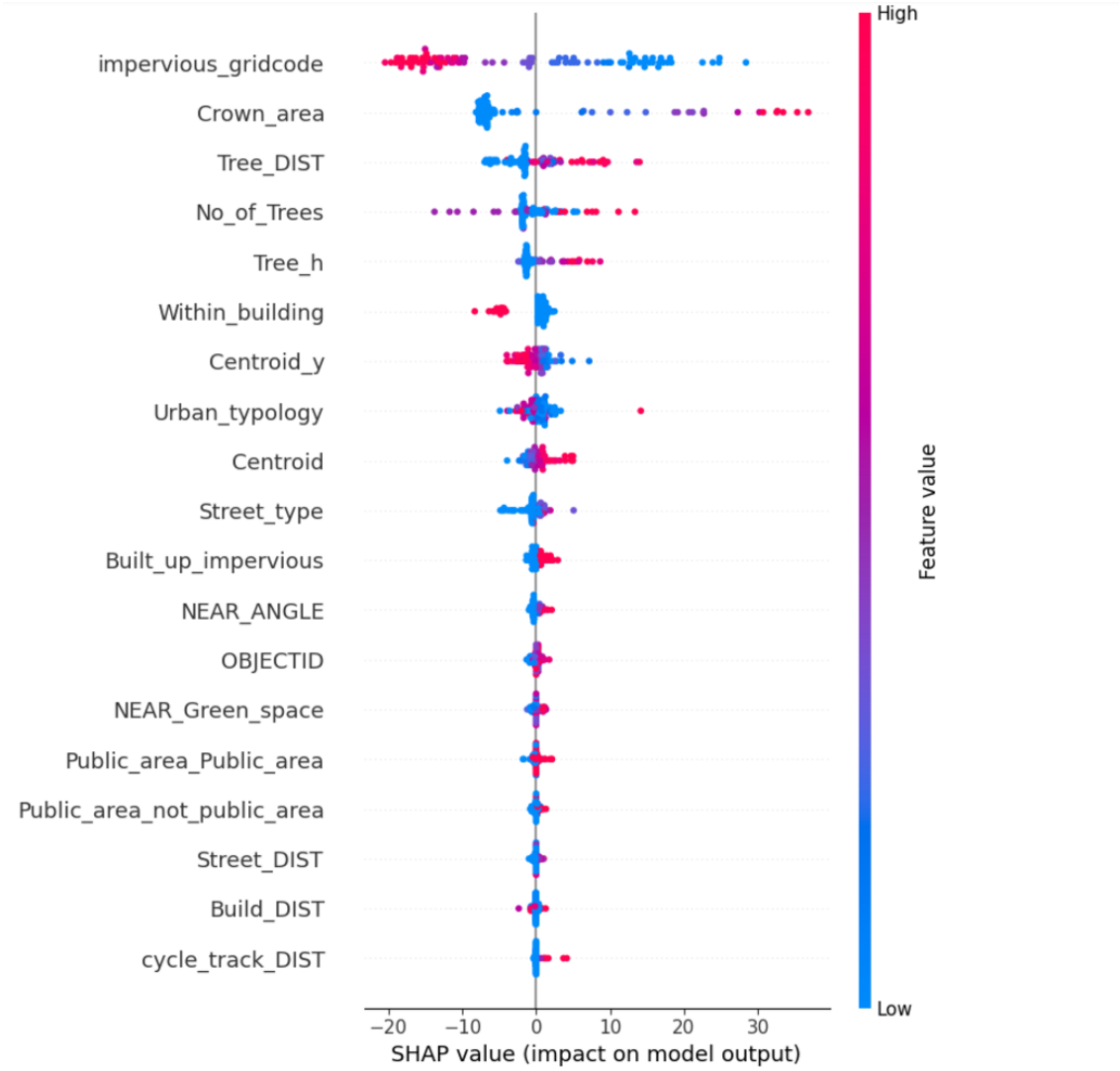

Figure 17 | Bee-swarm plot of SHAP values of all features showing an overview of prediction derived for model trained for condition 2 dataset of the main study

Figure 17 shows the bee swarm plot of the SHAP values for condition 2 of the main study, indicating the issue of multiple subsections for the categorical, like Urban typology and street type features, is now resolved, indicating the flexibility in other areas due to the maintenance consistent input shape read by the model for the dataset according to the framework. The Features like surface imperviousness (impervious_ grid code), vegetation data like crown area (Crown_area), distance to the nearest tree (Tree_Dist), built-up area imperviousness(Built_up_impervious) and building footprint (Within_building) are found to be the most significant factors influencing the predictions with the consistent distribution of SHAP values. The higher surface imperviousness values negatively impacted the greening potential, while the higher values of the crown area positively impacted the prediction.

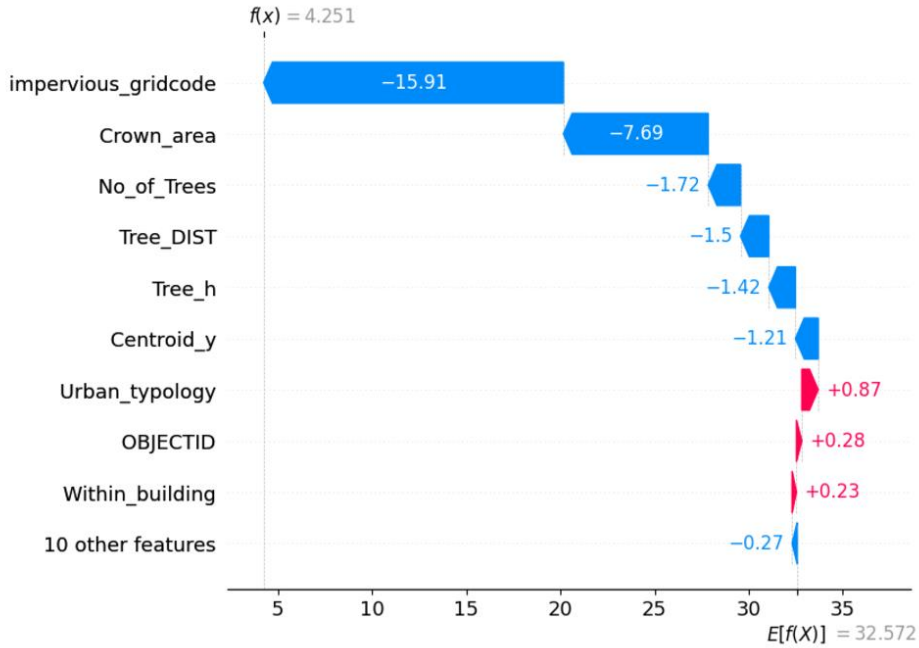

Figure 18 | Waterfall plot of SHAP values of a single sample of prediction derived for model trained for condition 2 dataset of the main study

Fig 18 illustrates the waterfall plot of the SHAP value of a single prediction, which exemplifies how the SHAP values of each contributing feature affect the prediction of the greening potential at that particular point. The analysis of the SHAP plot and the dataset of that particular point revealed that the point had higher imperviousness values and a low value of the crown area, along with other features, contributing to a lower greening potential.

The predicted test data of the model is extracted into CSV file format and exported into the GIS platform for further visualization and analysis. The subtraction difference between the predicted greening potential and actual greening potential values of each point is visualized in ArcGIS for interpretation (Refer to Figure 19).

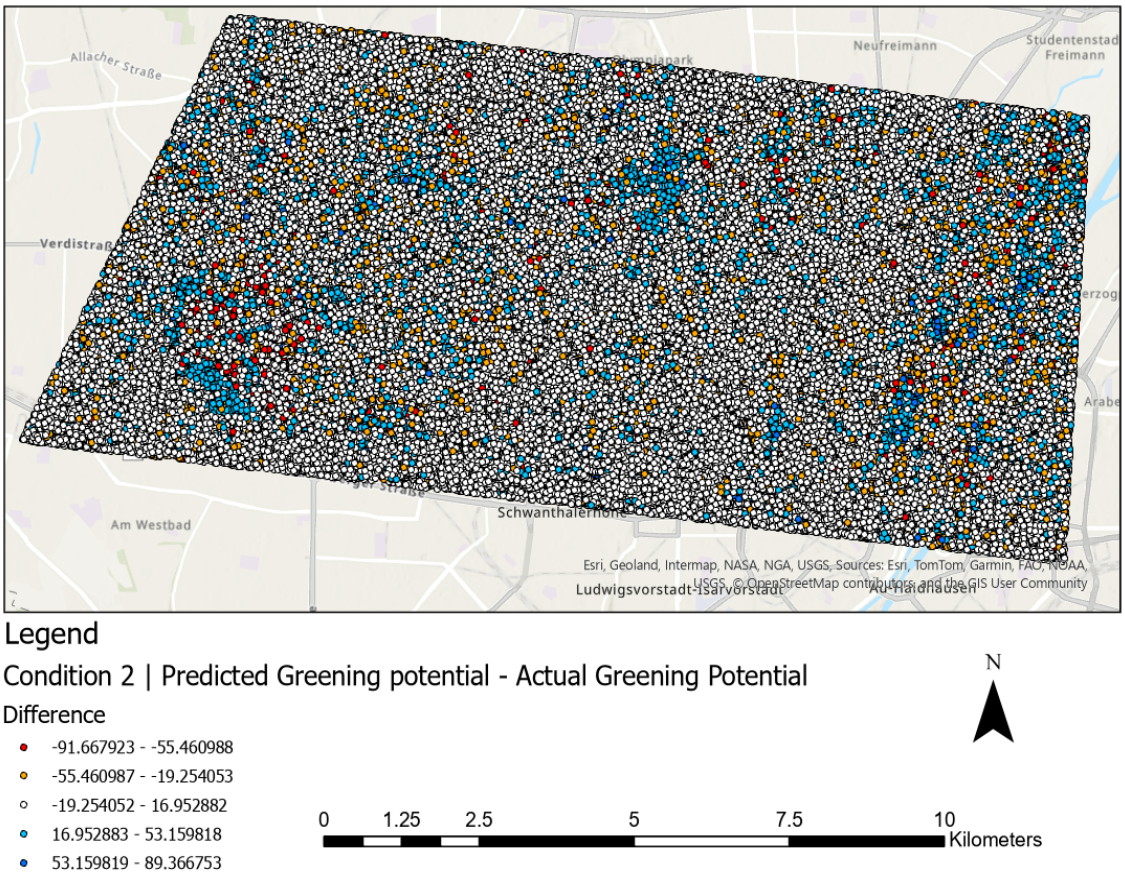

Figure 19 | Subtraction difference between the predicted Greening potential and Actual greening potential values of each point is visualized in ArcGIS for the model trained for Condition 2 of the main study

The test data prediction visualization helps identify locations with over-predictions and under-predictions for detailed analysis to optimise the model's performance. In Figure 19, the red colour points mark the points that underpredict the greening potential, and the dark blue represents the areas that are making overpredition. The light blue and orange points represent the area with slight overprediction and underprediction, respectively. The white points mark the locations where the model could make predictions aligned with the actual greening potential values. The overprediction and underprediction areas were mostly found near the "open or green spaces" and "forestry" land typologies, which can be looked upon for further optimization efforts through a thorough investigation with broader datasets.

5.2.3 Results from condition 3 model training

Model Training performance

This NN model was also trained with advanced optimization techniques to avoid overfitting and was configured to monitor the validation loss. The condition 3 model takes all the advantages of the condition 2 model to further optimization efforts. The training performance shows that the “final training loss” achieved was 233.58, and the “validation loss” was 229.68. The “validation loss” and “training loss” converge steadily, the model has better architecture than the rest of the conditions, and the model is the learning data better than the rest. The “final training loss” and the “validation loss” are remarkably lower than the rest of the conditions, and this model is also flexible. Table 10 shows the model training results of the condition 3 model.

Model Testing Evaluation

The model was evaluated on the test set, and the R^2 score given for the testing was 0.7916. This indicates that the model was able to explain approximately 79.16% of the variance in the target variable and better performance among all the other models, indicating better alignment of the prediction data with actual values (greening potential). This result suggests improvement in the overall model building and dataset preparation. The “test loss” MSE value was 228.68, and convergence with validation loss suggests better performance of the model's architecture.

Table 10 | Model Training Results for Main Study, condition 3

	Training Loss	**Validation Loss**	**Test Loss (MSE)**	**R^2 Score**
Output value	233.58	229.68	228.68	0.7916

Interpretation of the Test Results with Plot Diagram

The SHAP values are derived for the model training for data exploration, debugging and explanation. A bee-swarm plot was derived, which gives an overview of all predictions with all the SHAP value features. The waterfall plot of a particular model prediction was also derived to interpret the feature dependencies of the model training.

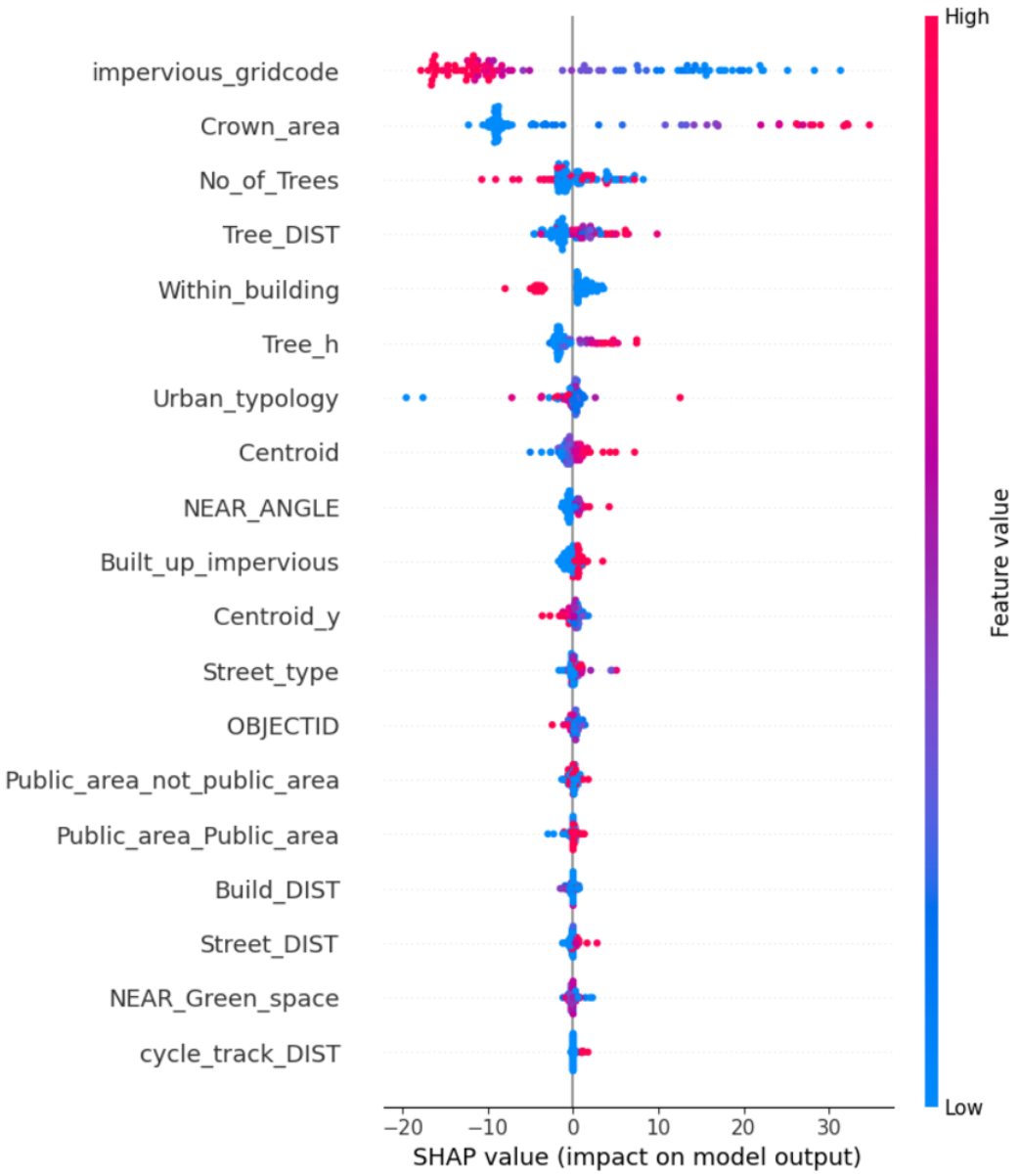

Figure 20 | Bee-swarm plot of SHAP values of all features showing an overview of prediction derived for model trained for condition 3 dataset of the main study

Figure 20 shows the bee swarm plot of the SHAP values for the condition 3 model of the main study, explaining the advantages of the condition 2 model and the better generalizability of the model. The Features like surface imperviousness (impervious_ grid code), vegetation data like crown area (Crown_area), distance to the nearest tree (Tree_Dist), proximity to near green space (NEAR_Green_space), built-up area imperviousness(Built_up_impervious) and building footprint (Within_building) are found to be the most significant factors influencing the predictions with the consistent distribution of SHAP values. The higher surface imperviousness values negatively impacted the greening potential, while the higher values of the crown area positively

impacted the prediction. The urban typology features' SHAP value distribution also showed better consistency in the distribution of the feature values for making predictions compared to the other 2 bee swarm plots of other models. Also, the lower values of the proximity to green spaces (NEAR_Green_space) positively impacted the greening potential predictions, suggesting that the greening potential prediction is positively impacted at locations closer to green spaces.

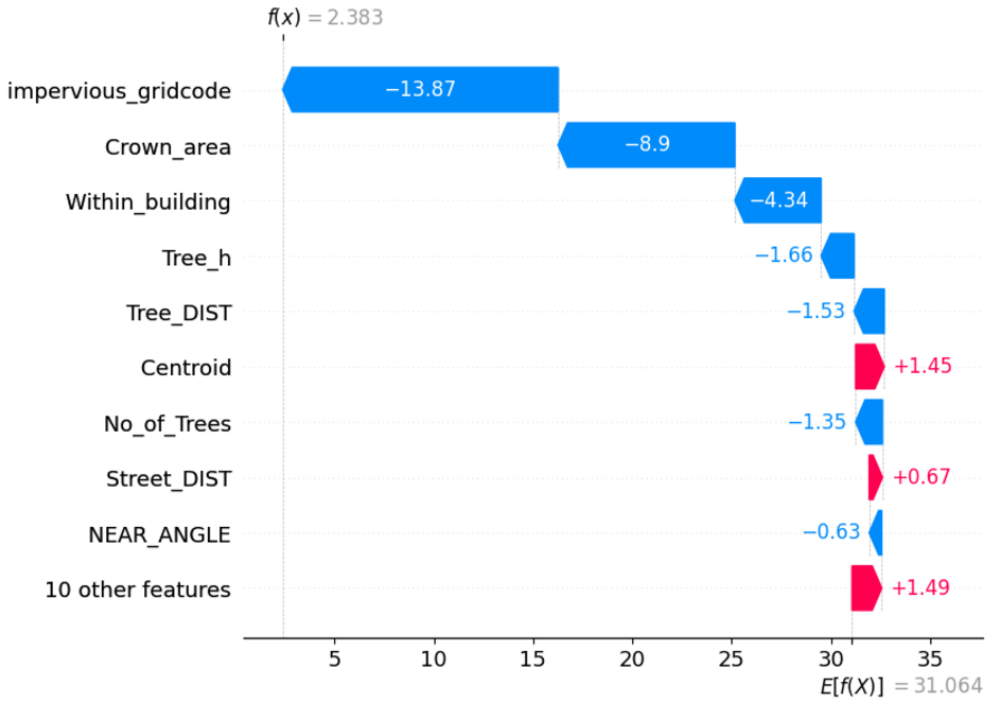

Figure 21 | Waterfall plot of SHAP values of a single sample of prediction derived for model trained for condition 3 dataset of the main study

Fig 21 illustrates the waterfall plot of the SHAP value of a single prediction, which exemplifies how the SHAP values of each contributing feature affect the prediction of the greening potential at that particular point. The analysis of the SHAP plot and the dataset of that particular point revealed that the point had higher imperviousness values, a low value of the crown area, and features suggesting that the point is within the building footprint, resulting in the prediction of lower greening potential at that point.

The predicted test data of the model is extracted into CSV file format and exported into the GIS platform for further visualization and analysis. The subtraction difference between the predicted greening potential and actual greening potential values of each point is visualized in ArcGIS for interpretation (Refer to Figure 22).

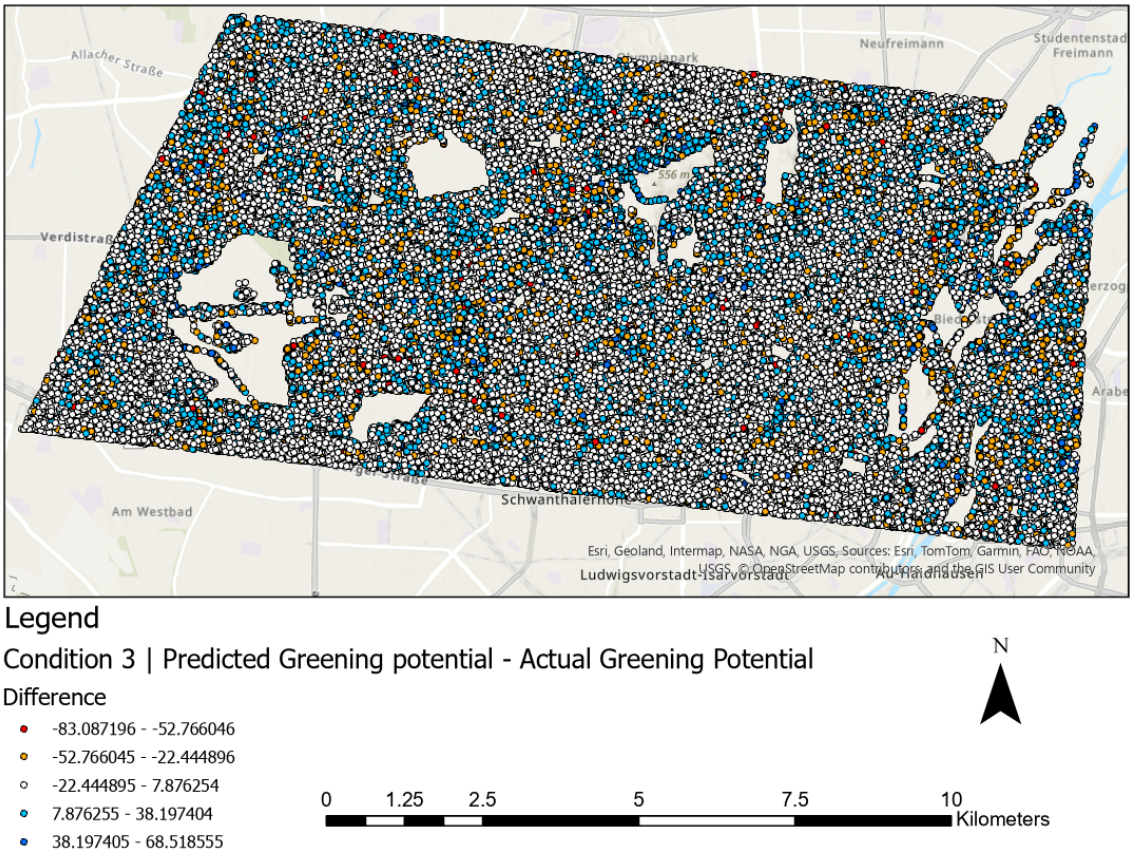

Figure 22 | Subtraction difference between the predicted Greening potential and Actual greening potential values of each point is visualized in ArcGIS for the model trained for Condition 3 of the main study

The test data prediction visualization helps identify locations with over-predictions and under-predictions for detailed analysis to optimise the model's performance. In Figure 22, the red colour points mark the points that underpredict the greening potential, and the dark blue represents the areas that are making overpredition. The light blue and orange points represent the area with slight overprediction and underprediction, respectively. The white points mark the locations where the model could make predictions aligned with the actual greening potential values. The removal of data in the areas of land typology, "open or green spaces", and "forestry" for model training resulted in the model better generalizing the data to make predictions. The areas with over-predictions and under-predictions also decreased. Further optimization of the locations with overprediction and underprediction can be looked at through a thorough investigation with broader datasets, which is scope for further studies.

5.3 Results from the Validation Study

The models trained based on Condtion 2 and Condtion 3 of the main study could be used in testing the model at the validation site as these models were flexible. The model trained for condition 2 is tested on the dataset prepared from the validation study to make predictions on the greening potential, and the resulting R2 score for predicting the data was 0.385. The lower value suggests the overfitting of the model, which suggests more optimization.

The model trained for condition 3 is tested on the dataset prepared from the same validation study to make predictions on the greening potential, and the resulting R^2 score for predicting the data was 0.485. The R^2 score improved, but the model still needs to be improved, and the overfitting of the data needs to be improved to reduce overfitting.

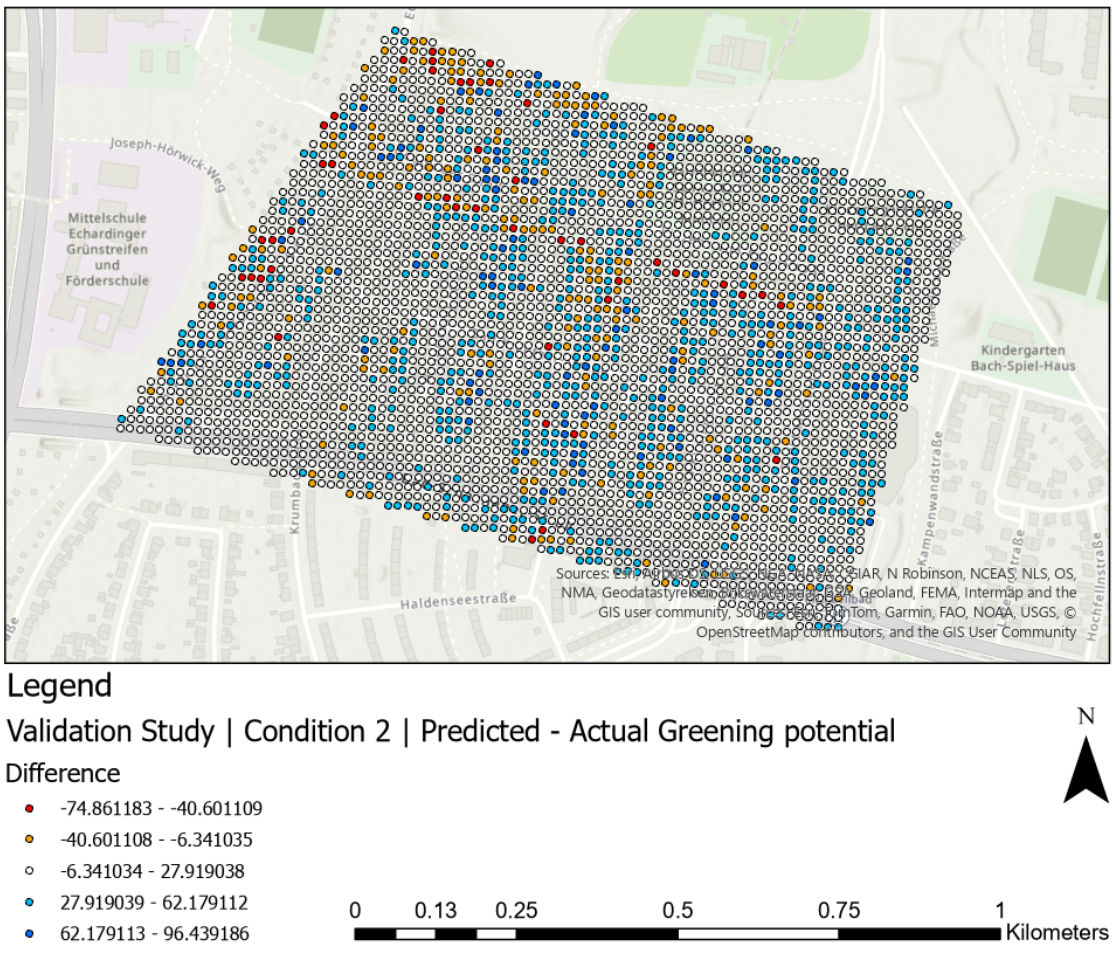

Figure 23 | Subtraction difference between the predicted Greening potential and Actual greening potential values of each point is visualized in ArcGIS. Prediction from the model on validation dataset, trained with Condition 2 of the main study

The prediction data of both models is visualized in the GIS platform. The prediction values are compared with that location's actual greening potential values. Fig 23 represents the subtraction difference between the predicted Greening potential and Actual greening potential values of each point visualized in ArcGIS. Predictions were derived from the model on the validation dataset and trained with condition 2 of the main study. In Figures 23 and 24, the red colour points mark the points that underpredict the greening potential, and the dark blue represents the areas that are making overpredictions. The light blue and orange points represent the area with slight overprediction and underprediction, respectively. The white points mark the locations where the model could make predictions aligned with the actual greening potential values. From Figure 23, we can infer that the model was able to predict many locations with precision as there were many white points, which indicates that the predictions of those locations were in line with the actual target variable values.

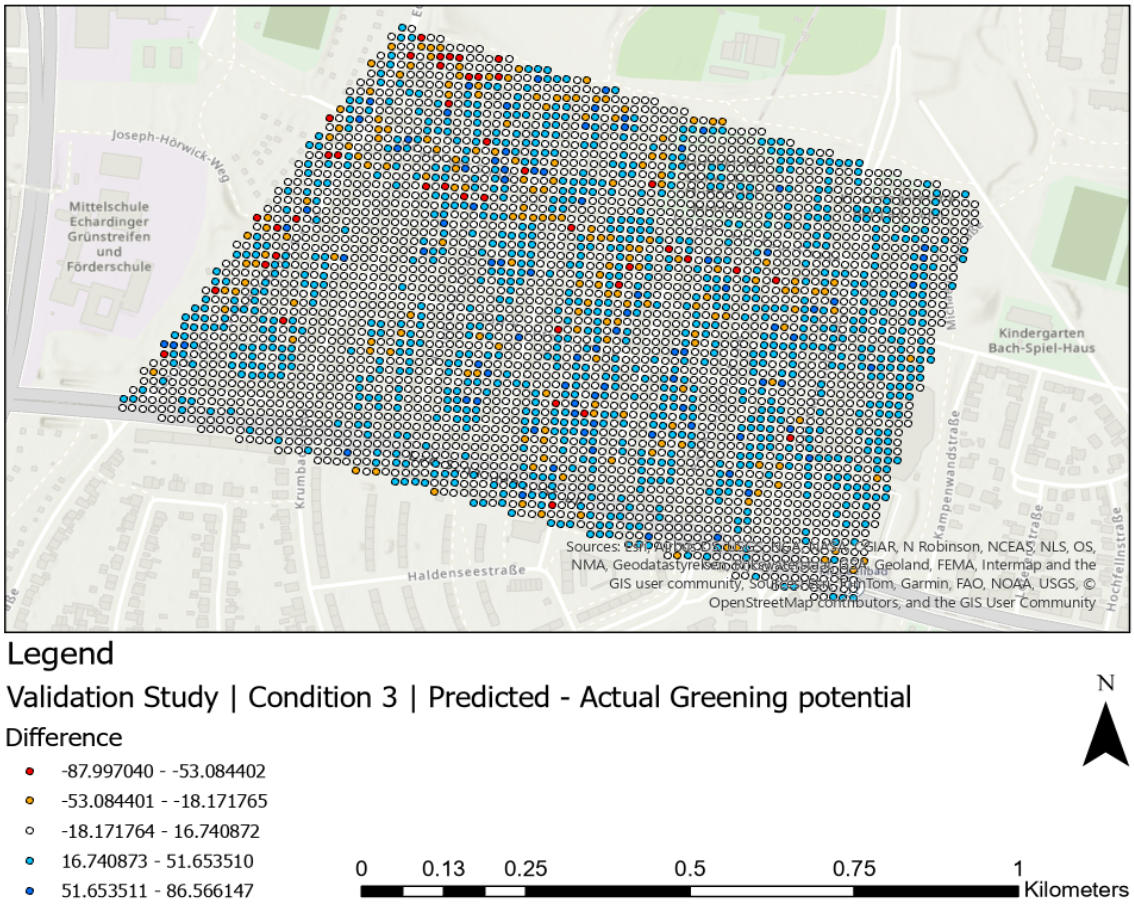

Figure 24 | Subtraction difference between the predicted Greening potential and Actual greening potential values of each point is visualized in ArcGIS. Predictions from the model on validation dataset, trained with Condition 3 of the main study

The Condition 2 model had comparatively more underprediction than the Condition 3 model in the validation study dataset. Fig 24 represents the subtraction difference between each point's predicted Greening potential and Actual greening potential values, visualized in ArcGIS. Predictions were derived from the model on the validation dataset and trained with Condition 3 of the main study. From Figure 24, we can infer that the model could predict many locations with precision as there were many white points, which indicates that the predictions of those locations were in line with the actual target variable values. Also, the model highlighted some locations with positive predictions where there was no existing vegetation. These overpredicted areas are marked in blue points, which can be further looked upon for potential locations for tree planting as these areas demonstrated properties for the better potential for tree planting such as lower surface imperviousness and land typologies like “linear building” which had better tree density as explained in table 3.

5.4 Results from Comparison of the model training using NN and Decision Tree Regression model

The dataset used in condition 1 of the main study is used for training the two models, and the prediction performance of the two model types is recorded in Table 11. The R^2 score and Test MSE values are better for NN, indicating its better performance and architecture.

Table 11 | Comparison results of model performance of NN and DecisionTree Regression model

	Neural Network Model	**Decision Tree Regression Model**
Test MSE	310.41	328.80
R^2 Score	0.7725	0.7590

6. Discussion

This study explored advanced machine learning techniques, including FNN, to predict the city's greening potential. The research focused on understanding predictive accuracy, feature importance, and the role of the city's various physical and spatial features contributing to greening potential. The methodology involved a rigorous pipeline of data preprocessing, model training, evaluation, and explainability analysis using SHAP values.

6.1 Comparison of different models

The NN model exhibited a lower MSE and higher R^2 score compared to the decision tree regression model, suggesting better predictive accuracy and superior architecture of the NN model. The training process of FNN included advanced optimization techniques such as dropout layers and early stopping, which likely contributed to its improved generalization and performance. The relatively low MSE indicates that the model's predictions are closer to the actual values, while the higher R^2 value reflects a better fit of the model, which enables making predictions which are comparatively better aligned with actual values. The better performance and architecture suggest the FNN model's suitability for this study, which encompassed the datasets containing complex feature-engineered data containing multiple non-linear relationship features.

The decision tree regression model had a slightly higher MSE and a lower R^2 score than the neural network. While the decision tree is a more straightforward and more interpretable model, it might have suffered from overfitting or could not capture the complexity of the relationships in the data. This explains the decision tree model’s struggle in generalizing large datasets involving complex relationships within data.

The FNN had advanced techniques, such as dropout layers and early stopping, which helped curb the model's overfitting. The results suggest that for the chosen model for study, FNN are better suited for the created structured feature-engineered dataset. This finding underscores the importance of the selection of the right model according to the datasets and objectives. The FNN’s ability to learn non-linear relationships, generalize complex data, advanced techniques to curb overfitting, scalability and predictive accuracy makes FNN better suited for this study.

With reference to the state-of-the-art section, ensemble models like Random Forest also performed for similar GIS-based studies. Their performance comparison with FNN can also be evaluated with the same dataset and feature engineering, which is a further scope of the study.

6.2 Framework for Model Creation

Several methodological aspects and references to similar studies have been dealt with to create a robust framework for the study, which focuses on combining GIS and ML to determine the target variables of the study. The methodological flow and division of various tasks established in the study helped assess the model's efficiency and scope for optimization as well as further data processing in intermediate stages. Data processing steps, including normalization and one-hot encoding of categorical features, ensured data quality and flexibility without compromising on the compatibility of data for model training. Multiple methodological aspects were considered in deriving the methodological flow chart of the tasks, which helped in the framework‘s reliability and adaptability for iterative tasks to improve the model. The feature engineering used for defining the target variable “greening potential” with respect to the crown projection area of the existing tree within the GIS environment was found to be a better practice, which reduces the load on human efforts, which further concretes the framework's stability. Tree canopy as an indicator was used in similar studies which focused on tree planting [62]. The methodological framework for defining spatial relationships and feature engineering the multiple features to make it compatible with model training has proven to be a solid approach, and the results of the study corroborate the effectiveness of the framework. This framework also provides a versatile approach, which can help in adjusting the model to perform diverse tasks as it leverages the scope for data scalability and flexibility.

The neural network’s architecture, incorporating multiple dense layers and dropout for regularization, demonstrated the ability to generalize well on unseen data. Additionally, early stopping functions were introduced to mitigate the overfitting of the model, improving model efficiency and robustness (Refer to Appendix 1 for the code used for model training). SHAP values enriched the study by making the predictions interpretable. The SHAP values interpretation of the features affecting the prediction of the model was in line with similar GIS-based studies. This is discussed in the coming section.

6.3 Creation of a flexible model based on improvement and validation

This study emphasizes the importance of iterative improvement and validation to develop a flexible machine-learning model capable of predicting greening potential effectively across diverse scenarios. The investigation involved three conditions which progressively optimized the model by modifying feature engineering and training methodologies. Each condition of the main study contributed unique insights. Condition 1 model training established a baseline by encoding categorical data using one-hot encoding, converting them into a format suitable for machine learning algorithms. Condition 2 sought to address the limitation of the baseline model by replacing one-hot encoding with numerical encoding for categorical features, thereby reducing the feature sparsity and enabling the model to generalize to regions with limited urban characteristics and scales. Condition 3 focused on avoiding features that could lead to overfitting and hinder the model's performance. Excluding land typologies like “open or green spaces” and “forestry” during model training was beneficial for improving the learning of the model. This step enhanced the performance of the model in generalizing the learning data and reducing the tendency of the model to learn noises in the data. So careful data assessment and feature engineering of the datasets according to the reviewed results can help create much more competent ML models for greening potential predictions.

This study emphasises the various challenges in the course of the study while implementing the model to make predictions on urban spatial data, which can be used as a roadmap for further studies in this field. This study also addresses the challenges faced, such as data sparsity, flexibility of the model in diverse conditions, overfitting of the model, and how these challenges are tackled through systematic approaches within the framework of the study. The validation of the model in entirely unseen data and its satisfactory performance underscores the objectives of this research study. This emphasises the transparency and trustworthiness of the model in the context of practical usage in urban planning.

6.4 Interpretation of the Feature importance SHAP in urban Tree planting prediction

Understanding the influence of individual independent variables/ features in making predictions on target variables by the model in regression tasks is crucial for improving and interpreting the performance and trustworthiness of the model. SHAP values are used to give insights into the significance of each independent contributing feature towards the prediction of the model in this study. The key features influencing the predictions in this study were the Crown area, surface imperviousness, number of nearby trees, distance or proximity to nearby green spaces and locations within building footprints.

The SHAP analysis of the prediction of the Condition 3 model in this study suggests that the large positive feature values crown area of the trees positively impact the greening potential. The higher crown area features indicate that space for expansion is not constrained and explain an area with existing UGI that needs proper maintenance. Similar observations were found in the study of urban tree crown development in urban environments where the positive impact of crown geometry of existing vegetation was influenced in that machine learning model [110].

The surface imperviousness feature also had a higher influence on the greening potential predictions as the SHAP plots suggest that the higher surface imperviousness values negatively impact the greening potential due to limited plantable area. Studies done by R Retiberger & et al. [39] suggested threshold values of surface imperviousness up to 81% were more suitable for tree planting, and above that limit, it may negatively affect tree planting possibilities. Another study done on combining multisource GIS methods and deep learning for spatial analyses in urban and greening changes found that increased built-up areas correlate with a decline in vegetation covers [111]. The within-building footprint feature is also similarly explained as the surface imperviousness feature.

Other major positively impacting features were the number of nearby trees and the proximity to existing green spaces. Their positive impact suggests the suitability for further planting of trees and emphasising the benefits of connected green spaces within urban areas. These observations aligned with the Urban tree planting suggestion and principles derived from the study by Rieke Hansen & et al. [22, 23]. A study using deep learning models on green view indices using Google Earth data found that proximity to

green spaces enhances the urban greenery assessments, which aligns with the findings of SHAP values of this study. These supporting studies and SHAP values interpretation of this study, validate the ML model's trustworthiness and robustness.

6.5 Data-Driven Decision-making

Integrating GIS and ML models in UGI studies can facilitate data-driven decision-making by urban planners and authorities, which can help optimise resource allocation and judicious planning for UGI. The model trained in this study is capable of identifying areas with higher greening potential by analyzing the significant spatial features like existing tree data and surface imperviousness data. The approach of using model predictions to identify locations which support high greening potential can ensure the resources are deployed to locations identified by the model, which can achieve the most significant impact, thus achieving resource efficiency in UGI development. Studies have shown that tree planting without a holistic approach has resulted in 50 per cent of the newly planted trees being lost within 5 years after planting due to above- and below-ground stressors [23, 40]. With this model prediction, urban areas can optimize the site selection and resource allocation for UGI development.

The SHAP analysis inferred that the model could explain the significance of existing vegetation and surface imperviousness features. This model and data framework can be used to identify locations with higher surface imperviousness and low vegetation, which can help in tree-planting efforts to mitigate UHI. A similar objective was used in a study that used GIS-based approaches to identify ideal tree planting locations along the streets, which incorporated multiple spatial and temporal data to plant trees to reduce temperature along streets[18, 32]. The validation study on the model revealed that the model was able to predict locations which have a higher greening potential with low vegetation cover. From the visualization of the prediction data in the validation study, it was observed that regions categorized as public property and "residential and linear building" land use showed positive trends in greening potential. Referring to Figure 24, in the validation study section, the locations with overprediction can also be overlooked and utilized for new tree planting locations. This emphasises a new possibility of the model to get an overview of the locations for new greening potential opportunities. These findings can aid in studies done by Dexter & et al. [27], which prioritized tree planting sites based on goals such as the needs of the community and the suitability of the locations. The provision of the model output integrated with the GIS tools also helped in

the geospatial visualization of each location's greening potential, which further enhances the data-driven knowledge and decision-making for urban planners. The model explained and incorporated multiple objectives, thus enabling better decision-making for the holistic development of the UGI.

6.6 Adaptability of this study in other cities

Cities with the availability of similar GIS data can utilize a similar methodology and framework to create an AI model to predict the greening potential of the city's expanding areas and improve the UGI within its existing area. This model can be utilized but with some adjustments involving fine-tuning the weights or retraining the model to account for geographical and region-specific features. Despite these changes, the applicability of the architecture of the current model looks promising and robust, owing to the flexibility of the model. This is of a further scope of study and detailed research.

Urban planners can employ this framework and methodology established in this to replicate it in other cities, as it follows a versatile approach. The results of this study suggest that by leveraging the GIS-based above-ground data and ML techniques, we can create a flexible model which can predict the greening potential of multiple cities, and this can form a guiding tool for the sustainable development of the cities.

6.7 Limitations of the study

The study provides insights into the Greening potential of the city with an ML model trained on GIS-based, data, but there is still room for improvement of the model and methodology. The validation study results showed that the model is still overfitting and suggested further feature engineering with comprehensive data. The dataset's reliance on specific spatial and physical features limits generalizability to areas where such data might not be available. Also, the quality of the model predictions depends on the quality of the data source from which the dataset for training is derived. So, if there are any shortcomings or deviations in the data used for training, it can affect the model’s performance.

Neural networks demonstrated sensitivity to hyperparameters, requiring careful tuning to achieve optimal performance. The results and evaluation of the training process suggested overfitting the data as the Neural network training was stopped in fewer epochs with early stopping functions. Models trained with fewer epochs or inappropriate

dropout settings performed poorly, indicating the need for computational resources and expertise.

This model and study were studied and applied only in one city, and the features available for Munich are considered for the study. Hence, it limits the understanding of how the model will perform on other cities and how it perform for entirely different urban topographies of other cities. Also, the readily available and open-source data are considered for creating the datasets as other potential parameters influencing the tree planting locations or greening potential predictions such as socio-economic factors, participation of non-governmental organization, communities, demographic features, meteorological or soil data and other parameters that can influence tree planting possibilities are not considered in developing this model [29].

Physical validation, like on-ground verification of the model's prediction on various sites, was not done. On-site verification of the model's prediction with the actual scenario of the site is crucial for the practical applicability and feasibility of the model.

Model applications in another area outside of the training area also require the process of creating a structured dataset similar to the training dataset but without the target variable. This dataset creation demands knowledge and is time-consuming to prepare similar datasets.

7 Conclusion

This study underscores the successful integration of ML techniques, particularly FNN, with Geographic Information Systems (GIS) to predict urban greening potential. The study's main objective was to create a framework that enables automation using AI on GIS-based data to predict the greening potential or locations suitable for new tree planting, which was accomplished through the methodology of this study. The model was able to generalize the data from GIS based above ground data to predict the possible locations for new tree planting sites based on effective data processing and feature engineering. Through iterative model improvements and rigorous evaluations, we demonstrated that FNN outperformed other models, such as decision tree regression, in predictive accuracy, robustness, and flexibility. The advantage of iterative methods in improving the model performance and the flexibility of being applied in different locations suggests the robustness and effectiveness of the framework used in this study. The results illustrate how physical, spatial, and categorical features contribute to greening potential in urban environments, enabling informed decision-making for urban planners and tree managers.

The results and discussion on the model's performance and flexibility indicate the study's strategically structured methodology and framework. The division of various methodology tasks and assigning the working environments for various tasks helped create a solid framework for model training and troubleshooting. The data preparation part enabled flexibility in integrating multiple datasets in the GIS platform, and their feature engineering using Geoprocessing tools helped in deriving the preliminary structured datasets for model training. The simplicity of Microsoft Office in handling structured datasets gives the advantage of the framework in scope for troubleshooting and easy data processing to make the datasets compatible with data training. The ML model selection based on the perspective of literature review and datasets further helped in model optimization through advanced techniques in model training. The study demonstrated the benefits of replacing one-hot encoding with numerical encoding to make the model adaptable to areas with diverse or missing land typologies and street classifications. This flexibility is essential when applying the model to cities with varying urban landscapes. Regularization techniques and early stopping functions enhanced model generalization and prevented overfitting in the model training. Model prediction visualization and SHAP analysis in the methodology adopted helped the study to explain

the predictions and trustworthiness of the model trained. The methodology used in the study employs a systematic approach, and the results from the study suggest a robust framework for the study.

The FNN models trained under three conditions progressively improved in accuracy, with the Condition 3 model achieving the best performance of the study with an R^2 score of 0.7916 and a mean Squared Error of 228.68. This highlights the importance of refining feature engineering, avoiding overfitting by excluding certain land types and adopting numerical encoding for categorical data, which helped improve the performance and flexibility of the model on diverse datasets. Furthermore, the SHAP analysis provided valuable insights into the influence of various features, offering interpretability and transparency for urban planning applications. The main contributing features, as explained using SHAP, emphasize that the data on existing nearby vegetation data, surface imperviousness data, building data, and proximity of the location to nearby green spaces had crucial and consistent effects on the prediction results.

Urban planners can utilize the data-driven decision-making insights provided by this model to prioritize greening efforts in the study area. The model can guide interventions, such as an overview of the existing greening potential of the area and locations with the potential for further tree planting based on the predictions from the model. From the visualization of the prediction data in the validation study, it was observed that regions categorized under public property and residential zones, like land typologies “linear building,” showed positive trends in greening potential. Excluding land typologies like “open or green spaces” and “forestry” during model training of Condition-3 models was found beneficial for improving the performance of the model (the “open or green spaces” and “forestry” land typologies are easily accessible for the planners (justification for exclusion)). Planners should carefully assess data features to prevent the ML models from overfitting and remain broadly applicable across different areas. Planners can focus and decide on under-served areas to build and develop UGI greening initiatives in these areas while ensuring a distributed network of urban greenery across all socioeconomic zones. integrating GIS for spatial data visualization and interpretation gives an extra overview to identify precise greening locations.

The model trained in this study is flexible to the varying test conditions, and this approach can be used in diverse urban areas within the city boundary. The model and framework used in this study, along with further research, can be made compatible with other cities with similar urban landscapes and the availability of GIS-based data.

The study also emphasizes the challenges faced during the model training and suggests ways to handle these challenges to create a flexible model that can predict greening potential. The study also comments on the limitations of the study, like the sparsity of data, physical validation on-site, and inferences made based on the use of the model in only one city. The study also comments on the sensitivity of the NN model and the chances of the model learning unintended relationships, leading to overfitting and underfitting of the model. It also comments on another downside of this study, which is the preparation of the dataset used for testing the model in other parts, which is a time-consuming process.

This study underlines that Urban planners should use GIS platforms with ML models for detailed project planning and monitoring, which can save resources and time. Integration of advanced technologies like ML with GIS data can help in the drive for a sustainable city with a holistic view.

8 Outlook and Further Research

This study focuses on creating a framework and methodology to integrate ML models with GIS-based data to derive automated models that predict the city's greening potential. This study emphasizes the importance of this study for the use of such models in practical application in urban areas, which can support the drive for sustainable cities with better development and management of UGI. The study also comments on various further scopes of this study and related research fields, which can help urban planners make data-driven decisions to develop sustainable cities.

This study outlines the scopes for scaling these models and the framework to employ them in other cities and urban areas by using the localized data of that particular city or urban area. Using this similar approach on various cities can create cross-city comparisons of the model and approach for better inference and to make decisions on the practical employability of the model in various terrains. Another scope would be training the model on urban landscapes, which are developed for their best potential of greening / UGI and assessing the models' prediction on regions, which are underdeveloped for UGI.

The study emphasises the scope of incorporating additional features to enhance the model's performance and efficiency. The features which can influence tree planting/ greening potential include land surface temperatures, tree growth patterns, soil data, underground, demographic features, surface imperviousness data with better resolution, etc. These features can help further optimise the model to attain a higher R^2 score without overfitting the data. The model can further be optimized for real-life usage by adopting input from active players in the cities, such as citizens, stakeholders, or organizations who are directly involved in the efforts for tree planting and maintenance. Their needs and insights can further help in incorporating more localized data into the model to enhance the practical usage of the model's predictions.

Some urban-based ML studies using GIS data also suggest the performance of EnsembleML models like Random Forest and Gradient Boosting machines [64]. These models can also be trained with similar datasets used in this study to evaluate the performance of those models in comparison with the FNN model used in this study. This can create more insights into the ML models, which can better generalize the large datasets.

Research can be done in the scope for building APIs for integrating this model with GIS tools, which can help in much better utilization of the model with real-time prediction and visualizations. Deploying the model to GIS tools or any other visualization tools using API can help the planners identify the gaps in greening potential within the cities. As a step after detailed optimization of the model, it can be used in tandem with augmented Reality or Virtual Reality to simulate greening potentials for better interpretation and visualization of the potential locations for UGI development.

Further scope involves the physical validation of the predictions made by the model on-site with careful assessment of the various factors affecting greening potential on-site. Further optimization based on these studies can help the model to be deployed for practical and policy making purposes.

9 Acknowledgment

The author expresses profound gratitude to his supervisor, Dr.-ing Roland Reitberger, for his invaluable guidance and unwavering support, which were instrumental in completing this study. He also especially thank the Chair of Energy Efficient and Sustainable Design and Building for providing essential guidance and resources for this study. Special thanks are extended to Dr. Tobias Leichte of Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR) for sharing the crucial data regarding the trees that served as this study's primary backing data. Completing this research work would not have been possible without all their contributions.

The author also thanks his family for their invaluable support throughout this journey. He especially thanks his Father, Mr. K. C. Preman, and brother, Mr. Vishnu Prem, for their continuous encouragement and provision of resources to pursue his studies. He expresses his gratitude to his mother, Mrs Sini Preman, and sister, Mrs Suganya Radhakrishnan, for their moral support and guidance, which greatly helped him complete his studies.

References

[1] A. Géron, *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (Covid-19 collection). Beijing, Boston, Farnham, Sebastopol, Tokyo: O'Reilly, 2019.

[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning* (Adaptive computation and machine learning). Cambridge Massachusetts: The MIT Press, 2016.

[3] Paul Bolstad, *GIS fundamentals*, 2002. [Online]. Available: https://www.researchgate.net/publication/200043155_GIS_fundamentals

[4] K. Gallatin and C. Albon, *Machine learning with Python cookbook: Practical solutions from preprocessing to deep learning*. Sebastopol CA: O'Reilly Media Inc, 2023.

[5] L. Gimmler, "A Deep Dive into Predicting Urban Growth using ArcGIS and R," *Esri*, 15 Jan., 2021. Accessed: Dec. 25, 2024. [Online]. Available: https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/a-deep-dive-into-predicting-urban-growth-using-arcgis-and-r/?utm_source=chatgpt.com

[6] Encyclopedia Britannica. "City | Definition & History | Britannica." Accessed: Nov. 14, 2024. [Online]. Available: https://www.britannica.com/topic/city

[7] J. E. Kohlhase, "The new urban world 2050: perspectives, prospects and problems," *Regional Science Policy & Practice*, vol. 5, no. 2, pp. 153–166, 2013. doi: 10.1111/rsp3.12001. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1757780223006728

[8] R. Cardoso, A. Sobhani, and E. Meijers, "The cities we need: Towards an urbanism guided by human needs satisfaction," *Urban Studies*, vol. 59, no. 13, pp. 2638–2659, 2022. doi: 10.1177/00420980211045571. [Online]. Available: https://www.researchgate.net/publication/355144750_The_cities_we_need_Towards_an_urbanism_guided_by_human_needs_satisfaction

[9] Peter Trowbridge and Nina Lauren Bassuk, *Trees in the Urban Landscape*, 2004. [Online]. Available: https://www.researchgate.net/publication/324506858_Trees_in_the_Urban_Landscape

[10] D. J. Nowak and J. T. Walton, "Projected Urban Growth (2000–2050) and Its Estimated Impact on the US Forest Resource," *Journal of Forestry*, vol. 103, no. 8, pp. 383–389, 2005. doi: 10.1093/jof/103.8.383. [Online]. Available: https://www.researchgate.net/publication/233585948_Projected_Urban_Growth_2000-2050_and_Its_Estimated_Impact_on_the_US_Forest_Resource

[11] C. Boyko, R. Cooper, and N. Dunn, Eds. *Designing future cities for wellbeing*. New York: RoutledgeTaylor & Francis Group, 2021.

[12] Y. Depietri and T. McPhearson, "Integrating the Grey, Green, and Blue in Cities: Nature-Based Solutions for Climate Change Adaptation and Risk Reduction,"

Nature-based Solutions to Climate Change in Urban Areas: Linkages Between Science, Policy, and Practice, pp. 91–109, 2017. doi: 10.1007/978-3-319-56091-5_6. [Online]. Available: https://www.researchgate.net/publication/317236775_Integrating_the_Grey_Green_and_Blue_in_Cities_Nature-Based_Solutions_for_Climate_Change_Adaptation_and_Risk_Reduction

[13] J. Wang, W. Zhou, and M. Jiao, "Location matters: planting urban trees in the right places improves cooling," *Frontiers in Ecol & Environ*, vol. 20, no. 3, pp. 147–151, 2022. doi: 10.1002/fee.2455. [Online]. Available: https://esajournals.onlinelibrary.wiley.com/doi/full/10.1002/fee.2455?saml_referrer

[14] D. Zhou, S. Zhao, L. Zhang, and S. Liu, "Remotely sensed assessment of urbanization effects on vegetation phenology in China's 32 major cities," *Remote Sensing of Environment*, vol. 176, pp. 272–281, 2016. doi: 10.1016/j.rse.2016.02.010. [Online]. Available: https://www.researchgate.net/publication/294258032_Remotely_sensed_assessment_of_urbanization_effects_on_vegetation_phenology_in_China's_32_major_cities

[15] Y. Wang and H. Akbari, "The effects of street tree planting on Urban Heat Island mitigation in Montreal," *Sustainable Cities and Society*, vol. 27, pp. 122–128, 2016. doi: 10.1016/j.scs.2016.04.013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S221067071630066X

[16] A. Fouillet *et al.,* "Excess mortality related to the August 2003 heat wave in France," *International archives of occupational and environmental health*, early access. doi: 10.1007/s00420-006-0089-4.

[17] "THE 17 GOALS | Sustainable Development." Accessed: Dec. 6, 2024. [Online]. Available: https://sdgs.un.org/goals

[18] R. Sousa-Silva, E. Cameron, and A. Paquette, "Prioritizing Street Tree Planting Locations to Increase Benefits for All Citizens: Experience From Joliette, Canada," *Front. Ecol. Evol.*, vol. 9, 2021, Art. no. 716611, doi: 10.3389/fevo.2021.716611.

[19] D. Dodman and D. Satterthwaite, "Institutional Capacity, Climate Change Adaptation and the Urban Poor," *IDS Bulletin*, vol. 39, no. 4, pp. 67–74, 2008, doi: 10.1111/j.1759-5436.2008.tb00478.x.

[20] S. Cohen, *The sustainable city*. New York: Columbia University Press, 2018.

[21] Mohammed M. Al‐Humaiqani and Sami G. Al‐Ghamdi, "Integrating Green‐Blue‐Gray Infrastructure for Sustainable Urban Flood Risk Management: Enhancing Resilience and Advantages," in *Sustainable cities in a changing climate: Enhancing urban resilience for the future*, S. G. Al-Ghamdi, Ed., Hoboken NJ: Wiley, 2024, pp. 207–226. [Online]. Available: https://www.researchgate.net/publication/376593309_Integrating_Green-Blue-Gray_Infrastructure_for_Sustainable_Urban_Flood_Risk_Management_Enhancing_Resilience_and_Advantages

[22] Rieke Hansen *et al., Urban Green Infrastructure. A foundation of attractive and sustainable cities. Pointers for municipal practice*. Bundesamt für Naturschutz, 2018.

[23] Rieke Hansen *et al., Urban Green Infrastructure. A foundation of attractive and sustainable cities. Pointers for municipal practice*. Bundesamt für Naturschutz, 2018. [Online]. Available: https://www.researchgate.net/publication/325386874_Urban_Green_Infrastructure_A_foundation_of_attractive_and_sustainable_cities_Pointers_for_municipal_practice

[24] H. Akbari, "Shade trees reduce building energy use and CO2 emissions from power plants," *Environmental Pollution*, vol. 116 Suppl 1, S119-26, 2002. doi: 10.1016/S0269-7491(01)00264-0. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0269749101002640

[25] D.J. Nowak and J.F. Dwyer, *Understanding the benefits and costs of urban forest ecosystems*, 2007. [Online]. Available: https://www.researchgate.net/publication/312470680_Understanding_the_benefits_and_costs_of_urban_forest_ecosystems

[26] Dexter H. Locke, J. Morgan Grove, Jacqueline W.T. Lu, Austin Troy, Jarlath P.M. O'Neil-Dunne, and Brian D. Beck, "Prioritizing Preferable Locations for Increasing Urban Tree Canopy in New York City,"

[27] D.J. Nowak and J.F. Dwyer, *Understanding the benefits and costs of urban forest ecosystems*, 2007. [Online]. Available: https://www.researchgate.net/publication/312470680_Understanding_the_benefits_and_costs_of_urban_forest_ecosystems

[28] D. J. Nowak, S. Hirabayashi, M. Doyle, M. McGovern, and J. Pasher, "Air pollution removal by urban forests in Canada and its effect on air quality and human health," *Urban Forestry & Urban Greening*, vol. 29, pp. 40–48, 2018. doi: 10.1016/j.ufug.2017.10.019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1618866717302182

[29] S. Pincetl, T. Gillespie, D. E. Pataki, S. Saatchi, and J.-D. Saphores, "Urban tree planting programs, function or fashion? Los Angeles and urban tree planting campaigns," *GeoJournal*, vol. 78, no. 3, pp. 475–493, 2013. doi: 10.1007/s10708-012-9446-x. [Online]. Available: https://link.springer.com/article/10.1007/s10708-012-9446-x

[30] F. E. Kuo and W. C. Sullivan, "Environment and Crime in the Inner City: Does Vegetation Reduce Crime?," *0000-0000*, vol. 33, no. 3, pp. 343–367, 2001. doi: 10.1177/00139160121973025. [Online]. Available: https://www.researchgate.net/publication/249624302_Environment_and_Crime_in_the_Inner_City_Does_Vegetation_Reduce_Crime

[31] N. Grima, W. Corcoran, C. Hill-James, B. Langton, H. Sommer, and B. Fisher, "The importance of urban natural areas and urban ecosystem services during the COVID-19 pandemic," *PLOS ONE*, vol. 15, no. 12, pp. 1–13, 2020. [Online]. Available: https://ideas.repec.org/a/plo/pone00/0243344.html

[32] R. Sousa-Silva, M. Duflos, C. Ordóñez Barona, and A. Paquette, "Keys to better planning and integrating urban tree planting initiatives," *Landscape and Urban Planning*, vol. 231, p. 104649, 2023. doi: 10.1016/j.landurbplan.2022.104649. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0169204622002985

[33] L. A. Roman, J. J. Battles, and J. R. McBride, "Urban tree mortality: a primer on demographic approaches," *Gen. Tech. Rep. NRS-158. Newtown Square, PA: U.S. Department of Agriculture, Forest Service, Northern Research Station. 24 p.*, vol. 158, pp. 1–24, 2016. doi: 10.2737/NRS-GTR-158. [Online]. Available: https://research.fs.usda.gov/treesearch/50688

[34] M. Schrenk *et al., Eds. REAL CORP 2021: Cities 20.50, creating habitats for the 3rd millennium, smart - sustainable - climate neutral: Proceedings of 26th International Conference on Urban Planning, Regional Development and Information Society = Beiträge zur 26. internationalen Konferenz zu Stadtplanung, Regionalentwicklung und Informationsgesellschaft*. Vienna: CORP - Competence Center of Urban and Regional Planning, 2021. [Online]. Available: https://archive.corp.at/cdrom2021/files/CORP2021_proceedings.pdf

[35] C. Wu, Q. Xiao, and E. G. McPherson, "A method for locating potential tree-planting sites in urban areas: A case study of Los Angeles, USA," *Urban Forestry & Urban Greening*, vol. 7, no. 2, pp. 65–76, 2008, doi: 10.1016/j.ufug.2008.01.002.

[36] D. Hilbert, L. Roman, A. Koeser, J. Vogt, and N. van Doorn, "Urban Tree Mortality: A Literature Review," *AUF*, vol. 45, no. 5, 2019, doi: 10.48044/jauf.2019.015.

[37] S. L. Harlan, A. J. Brazel, L. Prashad, W. L. Stefanov, and L. Larsen, "Neighborhood microclimates and vulnerability to heat stress," *Social Science & Medicine*, early access. doi: 10.1016/j.socscimed.2006.07.030.

[38] Xueqiao Huang and John R. Jensen, "A Machine-Learning Approach to Automated Knowledge-Base Building for Remote Sensing Image Analysis with GIS Data," [Online]. Available: https://www.asprs.org/wp-content/uploads/pers/1997journal/oct/1997_oct_1185-1194.pdf

[39] R. Reitberger, N. Pattnaik, and Y. Lu, Eds., *Urban tree placement analysis: a GIS-based approach for identifying suitable planting locations in Munich* (Technische Universität Hamburg, Institut für Digitales und Autonomes Bauen), 2024, doi: 10.15480/882.13517.

[40] N. Cavender and G. Donnelly, "Intersecting urban forestry and botanical gardens to address big challenges for healthier trees, people, and cities," *Plants People Planet*, vol. 1, no. 4, pp. 315–322, 2019. doi: 10.1002/ppp3.38. [Online]. Available: https://www.researchgate.net/publication/334303821_Intersecting_urban_forestry_and_botanical_gardens_to_address_big_challenges_for_healthier_trees_people_and_cities

[41] Hilbert and D. R, "Urban Tree Mortality: A Literature Review," [Online]. Available: https://www.fs.usda.gov/nrs/pubs/jrnl/2019/nrs_2019_hilbert_001.pdf

[42] E. W. Bodnaruk, C. N. Kroll, Y. Yang, S. Hirabayashi, D. J. Nowak, and T. A. Endreny, "Where to plant urban trees? A spatially explicit methodology to explore ecosystem service tradeoffs," *Landscape and Urban Planning*, vol. 157, pp. 457–467, 2017. doi: 10.1016/j.landurbplan.2016.08.016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016920461630175X

[43] Z. Cimburova, S. Blumentrath, and D. N. Barton, "Making trees visible: A GIS method and tool for modelling visibility in the valuation of urban trees," *Urban Forestry & Urban Greening*, vol. 81, p. 127839, 2023. doi: 10.1016/j.ufug.2023.127839. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1618866723000109

[44] D. Tiede, *A full GIS-Based workflow for tree identification and tree crown delineation using laser scanning* (XXXVI), 2005. [Online]. Available: https://www.researchgate.net/publication/216266454_A_full_GIS-Based_workflow_for_tree_identification_and_tree_crown_delineation_using_laser_scanning

[45] M. M. Isa and N. Othman, "Using Geographic Information System for Trees Assessment at Public Park," *Procedia - Social and Behavioral Sciences*, vol. 42, pp. 248–258, 2012. doi: 10.1016/j.sbspro.2012.04.188. [Online]. Available: https://www.researchgate.net/publication/257716000_Using_Geographic_Information_System_for_Trees_Assessment_at_Public_Park

[46] Ruben de Laat and Léon van Berlo, "Integration of BIM and GIS: The development of the CityGML GeoBIM extension," in *Advances in 3D geo-information sciences*, T. H. Kolbe, G. König, and C. Nagel, Eds., Berlin: Springer, 2011, pp. 211–225. [Online]. Available: https://www.researchgate.net/publication/226107766_Integration_of_BIM_and_GIS_The_development_of_the_CityGML_GeoBIM_extension

[47] Y. Deng, J. C. Cheng, and C. Anumba, "Mapping between BIM and 3D GIS in different levels of detail using schema mediation and instance comparison," *Automation in Construction*, vol. 67, pp. 1–21, 2016, doi: 10.1016/j.autcon.2016.03.006.

[48] M. Dwyer and R. Miller, "Using Gis to Assess Urban Tree Canopy Benefits and Surrounding Greenspace Distributions," *Arboriculture & Urban Forestry (AUF)*, vol. 25, no. 2, pp. 102–107, 1999. doi: 10.48044/jauf.1999.016. [Online]. Available: https://auf.isa-arbor.com/content/25/2/102

[49] N. Tohidi and R. B. Rustamov, "A Review of the Machine Learning in GIS for Megacities Application," in *Geographic Information Systems in Geospatial Intelligence*, R. B. Rustamov, Ed., IntechOpen, 2020.

[50] O. Hamdy, H. Gaber, M. S. Abdalzaher, and M. Elhadidy, "Identifying Exposure of Urban Area to Certain Seismic Hazard Using Machine Learning and GIS: A Case Study of Greater Cairo," *Sustainability*, vol. 14, no. 17, p. 10722, 2022. doi: 10.3390/su141710722. [Online]. Available: https://www.mdpi.com/2071-1050/14/17/10722

[51] G.E. McPherson, D. Nowak, and G. Heisler, "Quantifying urban forest structure, function, and value: The Chicago urban forest climate project," *Bulletin of the*

Ecological Society of America, vol. 76, no. 2, 1995. [Online]. Available: https://www.researchgate.net/publication/255946320_Quantifying_urban_forest_structure_function_and_value_The_Chicago_urban_forest_climate_project

[52] T. Varol, S. Gormus, S. Cengiz, H. B. Ozel, and M. Cetin, "Determining potential planting areas in urban regions," *Environ Monit Assess*, early access. doi: 10.1007/s10661-019-7299-1.

[53] R. W. Miller, R. J. Hauer, and L. P. Werner, *Urban forestry: Planning and managing urban greenspaces,* 3rd ed. Long Grove: Waveland, 2015.

[54] R. Sousa-Silva, M. Duflos, C. Ordóñez Barona, and A. Paquette, "Keys to better planning and integrating urban tree planting initiatives," *Landscape and Urban Planning*, vol. 231, p. 104649, 2023, doi: 10.1016/j.landurbplan.2022.104649.

[55] Y. Sun *et al.,* "Using machine learning to examine street green space types at a high spatial resolution: Application in Los Angeles County on socioeconomic disparities in exposure," *The Science of the total environment*, early access. doi: 10.1016/j.scitotenv.2021.147653.

[56] S. Tu and M. Zhang, "The Big Data Model for Urban Road Land Use Planning Is Based on a Neural Network Algorithm," *Computational Intelligence and Neuroscience*, early access. doi: 10.1155/2022/2727512.

[57] A. Moser, M. A. Rahman, H. Pretzsch, S. Pauleit, and T. Rötzer, "Inter- and intraannual growth patterns of urban small-leaved lime (Tilia cordata mill.) at two public squares with contrasting microclimatic conditions," *International journal of biometeorology*, early access. doi: 10.1007/s00484-016-1290-0.

[58] T. Rötzer, M. A. Rahman, A. Moser-Reischl, S. Pauleit, and H. Pretzsch, "Process based simulation of tree growth and ecosystem services of urban trees under present and future climate conditions," *The Science of the total environment*, early access. doi: 10.1016/j.scitotenv.2019.04.235.

[59] M. B. Anteneh, D. S. Damte, S. G. Abate, and A. A. Gedefaw, "Geospatial assessment of urban green space using multi-criteria decision analysis in Debre Markos City, Ethiopia," *Environ Syst Res*, vol. 12, no. 1, pp. 1–21, 2023. doi: 10.1186/s40068-023-00291-x. [Online]. Available: https://environmentalsystemsresearch.springeropen.com/articles/10.1186/s40068-023-00291-x?utm_source=chatgpt.com

[60] A. Alhammad, Q. Sun, and Y. Tao, "Optimal Solar Plant Site Identification Using GIS and Remote Sensing: Framework and Case Study," *Energies*, vol. 15, no. 1, p. 312, 2022, doi: 10.3390/en15010312.

[61] M. C. Dwyer and R. W. Miller, "Using GIS to Assess Urban Tree Canopy Benefits and Surrounding Greenspace Distributions," *Arboriculture & Urban Forestry (AUF)*, vol. 25, no. 2, pp. 102–107, 1999. doi: 10.48044/jauf.1999.016. [Online]. Available: https://auf.isa-arbor.com/content/25/2/102?utm_source=chatgpt.com

[62] A. Morani, D. J. Nowak, S. Hirabayashi, and C. Calfapietra, "How to select the best tree planting locations to enhance air pollution removal in the

MillionTreesNYC initiative," *Environmental pollution (Barking, Essex : 1987)*, early access. doi: 10.1016/j.envpol.2010.11.022.

[63] C. V. Ekeanyanwu, I. F. Obisakin, P. Aduwenye, and N. Dede-Bamfo, "Merging GIS and Machine Learning Techniques: A Paper Review," *GEP*, vol. 10, no. 09, pp. 61–83, 2022. doi: 10.4236/gep.2022.109004. [Online]. Available: https://www.scirp.org/journal/paperinformation?paperid=119857

[64] Joan M. Peralta and Thelma D. Palaoag, "Exploring Machine Learning for Urban Green Space Mapping in Smart Cities: A Scoping Review," *NANO-NTP*, 193–205-193–205, 2024. doi: 10.62441/nano-ntp.vi.1261. [Online]. Available: https://nano-ntp.com/index.php/nano/article/view/1261

[65] J. Rustamov, Z. Rustamov, and N. Zaki, "Green Space Quality Analysis Using Machine Learning Approaches," *Sustainability*, vol. 15, no. 10, p. 7782, 2023. doi: 10.3390/su15107782. [Online]. Available: https://www.mdpi.com/2071-1050/15/10/7782

[66] Siddhant Ray, *A Comparative Analysis and Testing of Supervised Machine Learning Algorithms*, 2018. [Online]. Available: https://www.researchgate.net/publication/335243082_A_Comparative_Analysis_and_Testing_of_Supervised_Machine_Learning_Algorithms

[67] A. Gulli and S. Pal, *Deep Learning with Keras*. Packt Publishing Ltd, 2017.

[68] G. Zaccone and M. R. Karim, *Deep Learning with TensorFlow: Explore neural networks and build intelligent systems with Python, 2nd Edition*. Packt Publishing Ltd, 2018.

[69] X. Huang, J. R. Jensen, and Mackey, H. E., Jr., "A machine learning approach to automated construction of knowledge bases for expert systems for remote sensing image analysis with GIS data," Westinghouse Savannah River Company, Rep. CONF-9604133--1, Jun. 2015. [Online]. Available: https://digital.library.unt.edu/ark:/67531/metadc670345/

[70] K. Zhang and M. Chen, "Multi-method analysis of urban green space accessibility: Influences of land use, greenery types, and individual characteristics factors," *Urban Forestry & Urban Greening*, vol. 96, p. 128366, 2024. doi: 10.1016/j.ufug.2024.128366. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S161886672400164X

[71] S. K. Hanoon, A. F. Abdullah, H. Z. M. Shafri, and A. Wayayok, "Urban Growth Forecast Using Machine Learning Algorithms and GIS-Based Novel Techniques: A Case Study Focusing on Nasiriyah City, Southern Iraq," *IJGI*, vol. 12, no. 2, p. 76, 2023. doi: 10.3390/ijgi12020076. [Online]. Available: https://www.mdpi.com/2220-9964/12/2/76?utm_source=chatgpt.com

[72] Z. Huang, H. Qi, C. Kang, Y. Su, and Y. Liu, "An Ensemble Learning Approach for Urban Land Use Mapping Based on Remote Sensing Imagery and Social Sensing Data," *Remote Sensing*, vol. 12, no. 19, p. 3254, 2020. doi: 10.3390/rs12193254. [Online]. Available: https://www.mdpi.com/2072-4292/12/19/3254?utm_source=chatgpt.com

[73] S. S. Band *et al.,* "Novel Ensemble Approach of Deep Learning Neural Network (DLNN) Model and Particle Swarm Optimization (PSO) Algorithm for Prediction of Gully Erosion Susceptibility," *Sensors (Basel, Switzerland)*, early access. doi: 10.3390/s20195609.

[74] A. Jain, A. Fandango, and A. Kapoor, *TensorFlow Machine Learning Projects,* 1st ed. [Erscheinungsort nicht ermittelbar], Sebastopol, CA: Packt Publishing; O'Reilly Media Inc, 2018. [Online]. Available: https://books.google.de/books?id=4i59DwAAQBAJ

[75] X. X. Zhu *et al.,* "Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, 2017. doi: 10.1109/MGRS.2017.2762307. [Online]. Available: https://www.researchgate.net/publication/322278874_Deep_Learning_in_Remote_Sensing_A_Comprehensive_Review_and_List_of_Resources

[76] Paul Bolstad, *GIS fundamentals*, 2002. [Online]. Available: https://www.researchgate.net/publication/200043155_GIS_fundamentals

[77] B. Nikparvar and J.-C. Thill, "Machine Learning of Spatial Data," *IJGI*, vol. 10, no. 9, p. 600, 2021. doi: 10.3390/ijgi10090600. [Online]. Available: https://www.mdpi.com/2220-9964/10/9/600

[78] "Haining, R. (2009) The Special Nature of Spatial Data. In Fotheringham, A.S. and Rogerson, P.A., Eds., The SAGE Handbook of Spatial Analysis, SAGE Publications, London, 5-23. - References - Scientific Research Publishing." Accessed: Dec. 29, 2024. [Online]. Available: https://www.scirp.org/reference/referencespapers?referenceid=2173857

[79] M. F. Goodchild, W. Li, and D. Tong, "Introduction to the special issue on scale and spatial analytics," *J Geogr Syst*, vol. 24, no. 3, pp. 285–289, 2022. doi: 10.1007/s10109-022-00391-9. [Online]. Available: https://www.researchgate.net/publication/362319642_Introduction_to_the_special_issue_on_scale_and_spatial_analytics

[80] A. Abdollahi and B. Pradhan, "Urban Vegetation Mapping from Aerial Imagery Using Explainable AI (XAI)," *Sensors*, early access. doi: 10.3390/s21144738.

[81] A. Sen and K. Gumus, "Comparison of Different Parameters of Feedforward Backpropagation Neural Networks in DEM Height Estimation for Different Terrain Types and Point Distributions," *Systems*, vol. 11, no. 5, p. 261, 2023. doi: 10.3390/systems11050261. [Online]. Available: https://www.mdpi.com/2079-8954/11/5/261

[82] T. Zhang, L. Wang, Y. Hu, W. Zhang, and Y. Liu, "Measuring Urban Green Space Exposure Based on Street View Images and Machine Learning," *Forests*, vol. 15, no. 4, p. 655, 2024. doi: 10.3390/f15040655. [Online]. Available: https://www.mdpi.com/1999-4907/15/4/655?utm_source=chatgpt.com

[83] J. Yu, P. Zeng, Y. Yu, H. Yu, L. Huang, and D. Zhou, "A Combined Convolutional Neural Network for Urban Land-Use Classification with GIS Data," *Remote Sensing*, vol. 14, no. 5, p. 1128, 2022. doi: 10.3390/rs14051128. [Online]. Available: https://www.mdpi.com/2072-4292/14/5/1128

[84] K. J. Bergen, P. A. Johnson, Maarten V. de Hoop, and G. C. Beroza, "Machine learning for data-driven discovery in solid Earth geoscience," *American Association for the Advancement of Science*, 22 Mar., 2019. Accessed: Dec. 28, 2024. [Online]. Available: https://www.science.org/doi/10.1126/science.aau0323#sec-2

[85] R. Pires de Lima and K. Marfurt, "Convolutional Neural Network for Remote-Sensing Scene Classification: Transfer Learning Analysis," *Remote Sensing*, vol. 12, no. 1, p. 86, 2020. doi: 10.3390/rs12010086. [Online]. Available: https://www.researchgate.net/publication/338167783_Convolutional_Neural_Network_for_Remote-Sensing_Scene_Classification_Transfer_Learning_Analysis

[86] G. Zhang, M. Wang, and K. Liu, "Deep neural networks for global wildfire susceptibility modelling," *Ecological Indicators*, vol. 127, p. 107735, 2021. doi: 10.1016/j.ecolind.2021.107735. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X21004003

[87] M. Alkahtani, J. Mallick, S. Alqadhi, M. N. Sarif, M. Fatahalla Mohamed Ahmed, and H. G. Abdo, "Interpretation of Bayesian-optimized deep learning models for enhancing soil erosion susceptibility prediction and management: a case study of Eastern India," *Geocarto International*, vol. 39, no. 1, 2024, Art. no. 2367611, doi: 10.1080/10106049.2024.2367611.

[88] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015. doi: 10.1038/nature14539. [Online]. Available: https://www.nature.com/articles/nature14539

[89] X. R. Zhang and G. Chen, "Application of Neural Network in Urban Land Use Suitability Evaluation," *KEM*, vol. 474-476, pp. 681–686, 2011. doi: 10.4028/www.scientific.net/KEM.474-476.681. [Online]. Available: https://www.researchgate.net/publication/271999727_Application_of_Neural_Network_in_Urban_Land_Use_Suitability_Evaluation

[90] J. Liu and S. Xu, "Advancements of Graph Neural Networks in Urban Traffic Prediction," in *Proceedings of the 1st International Conference on Engineering Management, Information Technology and Intelligence*, Shanghai, China, 2024, pp. 62–66, doi: 10.5220/0012902200004508.

[91] M. Zygmunt and D. Gawin, "Application of Artificial Neural Networks in the Urban Building Energy Modelling of Polish Residential Building Stock," *Energies*, vol. 14, no. 24, p. 8285, 2021. doi: 10.3390/en14248285. [Online]. Available: https://www.mdpi.com/1996-1073/14/24/8285?utm_source=chatgpt.com

[92] "An introduction to explainable AI with Shapley values — SHAP latest documentation." Accessed: Dec. 30, 2024. [Online]. Available: https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

[93] N. Bao *et al.,* "Towards interpreting machine learning models for understanding the relationship between vegetation growth and climate factors: A case study of the Anhui Province, China," *Ecological Indicators*, vol. 167, p. 112636, 2024. doi:

10.1016/j.ecolind.2024.112636. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1470160X24010938

[94] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development," *Clinical and Translational Science*, vol. 17, no. 11, e70056, 2024. doi: 10.1111/cts.70056. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC11513550/?utm_source=chatgpt.com

[95] W. E. Marcilio and D. M. Eler, "From explanations to feature selection: assessing SHAP values as feature selection mechanism," in *2020 33rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI)*, Recife/Porto de Galinhas, Brazil, 2020, pp. 340–347, doi: 10.1109/SIBGRAPI51738.2020.00053.

[96] "Data formats supported by Spatial Analyst—ArcGIS Pro | Documentation." Accessed: Dec. 6, 2024. [Online]. Available: https://pro.arcgis.com/en/pro-app/latest/help/analysis/spatial-analyst/basics/data-formats-supported-by-spatial-analyst.htm

[97] S. Jochner, M. Alves‐Eigenheer, A. Menzel, and L. P. C. Morellato, "Using phenology to assess urban heat islands in tropical and temperate regions," *Intl Journal of Climatology*, vol. 33, no. 15, pp. 3141–3151, 2013. doi: 10.1002/joc.3651. [Online]. Available: https://www.researchgate.net/publication/257147291_Using_phenology_to_assess_urban_heat_islands_in_tropical_and_temperate_regions

[98] M. A. Rahman, A. Moser, T. Rötzer, and S. Pauleit, "Microclimatic differences and their influence on transpirational cooling of Tilia cordata in two contrasting street canyons in Munich, Germany," *Agricultural and Forest Meteorology*, vol. 232, no. 4, pp. 443–456, 2017. doi: 10.1016/j.agrformet.2016.10.006. [Online]. Available: https://www.researchgate.net/publication/308904264_Microclimatic_differences_and_their_influence_on_transpirational_cooling_of_Tilia_cordata_in_two_contrasting_street_canyons_in_Munich_Germany

[99] "Wetter und Klima - Deutscher Wetterdienst - Leistungen - Vieljährige Mittelwerte." Accessed: Dec. 31, 2024. [Online]. Available: https://www.dwd.de/DE/leistungen/klimadatendeutschland/vielj_mittelwerte.html

[100] M. Schrenk *et al., Eds. REAL CORP 2021: Cities 20.50, creating habitats for the 3rd millennium, smart - sustainable - climate neutral: Proceedings of 26th International Conference on Urban Planning, Regional Development and Information Society = Beiträge zur 26. internationalen Konferenz zu Stadtplanung, Regionalentwicklung und Informationsgesellschaft*. Vienna: CORP - Competence Center of Urban and Regional Planning, 2021. [Online]. Available: https://archive.corp.at/cdrom2021/files/CORP2021_proceedings.pdf

[101] S. Pauleit *et al.,* "Results Brochure of the Research Training Group Urban Green Infrastructure – Training Next Generation Professionals for Integrated Urban Planning Research," 2024, doi: 10.14459/2024MD1742953.

[102] M. Werner, H. Li, and Zollner Max, Teuscher, Balthasar, "Bavaria Building Dataset (BBD)," 2023, doi: 10.14459/2023MP1709451.

[103] "High Resolution Layer Imperviousness — Copernicus Land Monitoring Service." Accessed: Dec. 31, 2024. [Online]. Available: https://land.copernicus.eu/en/products/high-resolution-layer-imperviousness

[104] "LHM_Stadtteilstudie_Webversion_04," [Online]. Available: https://stadt.muenchen.de/dam/jcr:1b6d455d-798a-4947-ab57-b8d07d17c7f9/LHM_Stadtteilstudie_Webversion_04.pdf

[105] "Raddauerzählstellen - Open Data München." Accessed: Dec. 31, 2024. [Online]. Available: https://opendata.muenchen.de/ru/pages/raddauerzaehlstellen

[106] "overpass turbo." Accessed: Dec. 31, 2024. [Online]. Available: https://overpass-turbo.eu/

[107] L. München. "GeodatenService München." Accessed: Dec. 5, 2024. [Online]. Available: https://stadt.muenchen.de/infos/portrait-geodatenservice.html

[108] "Impervious Built-up 2018 (raster 10 m and 100 m), Europe, 3-yearly — Copernicus Land Monitoring Service." Accessed: Dec. 31, 2024. [Online]. Available: https://land.copernicus.eu/en/products/high-resolution-layer-imperviousness/impervious-built-up-2018#download

[109] L. München. "Nachhaltige Stadtentwicklung." Accessed: Jan. 2, 2025. [Online]. Available: https://stadt.muenchen.de/infos/nachhaltige-stadtentwicklung-muenchen.html

[110] H. Yazdi, A. Moser-Reischl, T. Rötzer, F. Petzold, and F. Ludwig, "Machine learning-based prediction of tree crown development in competitive urban environments," *Urban Forestry & Urban Greening*, vol. 101, no. 25, p. 128527, 2024. doi: 10.1016/j.ufug.2024.128527. [Online]. Available: https://www.researchgate.net/publication/384481808_Machine_Learning-Based_Prediction_of_Tree_Crown_Development_in_Competitive_Urban_Environments

[111] M. Francini, C. Salvo, and A. Vitale, "Combining Deep Learning and Multi-Source GIS Methods to Analyze Urban and Greening Changes," *Sensors*, early access. doi: 10.3390/s23083805.

List of Figures

List of Tables

Appendix 1 | Script Used for Study

Neural Network Model code used for Main Study

```
import pandas as pd
import numpy as np
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import Dense, Dropout, Input
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error, r2_score
import shap
import matplotlib.pyplot as plt
from google.colab import files  # Only if you're using Google Colab as IDE
from google.colab import drive

# Mount your Google Drive
drive.mount('/content/drive')

# Load the data
file_path = '/content/drive/MyDrive/Colab Notebooks/Main_study.xlsx'
data = pd.read_excel(file_path)

# Assuming `data` is your DataFrame
features = [
    'OBJECTID', 'Centroid', 'Centroid_y', 'No_of_Trees', 'Tree_DIST', 'Crown_area',
    'Tree_h',    'NEAR_ANGLE',    'Within_building',    'Build_DIST', 'Built_up_impervious',
```

```
    'Street_DIST',          'Street_type',          'cycle_track_DIST',
'impervious_gridcode',
    'Urban_typology', 'NEAR_Green_space',
]

# Categorical features (One-Hot Encoding)
categorical_features = ['Public_area']
data_encoded = pd.get_dummies(data[categorical_features])

# Normalize NEAR_ANGLE to be within [0, 360]
data['NEAR_ANGLE'] = data['NEAR_ANGLE'] % 360

# Combine numerical and categorical features
X = pd.concat([data[features], data_encoded], axis=1)
y = data['Greening_potential']

# Split training and testing sets
X_train,   X_test,   y_train,   y_test   =   train_test_split(X,   y,
test_size=0.2, random_state=42)
test_indices = X_test.index  # Store original test set indices

# Standardize the data
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Verify Within_building is in the dataset
assert 'Within_building' in X.columns, "Within_building is not in the
feature list!"

# Define the neural network model
model = Sequential([
    Input(shape=(X_train_scaled.shape[1],)),  # Explicitly define input
shape
```

```
    Dense(128, activation='relu'),
    Dropout(0.4),
    Dense(128, activation='relu'),
    Dropout(0.4),
    Dense(64, activation='relu'),
    Dense(1)  # Output layer for regression (1 output node)
])

# Compile the model
model.compile(optimizer='adam',            loss='mean_squared_error',
metrics=['mean_squared_error'])

# Early stopping callback to prevent overfitting
early_stopping = tf.keras.callbacks.EarlyStopping(monitor='val_loss',
patience=3)

# Train the model
model.fit(X_train_scaled,    y_train,    epochs=50,    batch_size=32,
validation_split=0.2, callbacks=[early_stopping])

# Evaluate the model on the test set
test_loss, test_mse = model.evaluate(X_test_scaled, y_test, verbose=0)
print(f'Test set loss: {test_loss}')

# Make predictions on the test set
y_pred = np.clip(model.predict(X_test_scaled), 0, 100)

# Calculate Mean Squared Error and R^2 Score
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R^2 Score: {r2}")

# Save the model
```

```
model.save('my_model_new.h5')


# Save the model and weights in Google Drive (if using Colab)
model.save('/content/drive/MyDrive/Colab Notebooks/my_model_new.h5')
model.save_weights('/content/drive/MyDrive/Colab
Notebooks/my_model_weights.h5')


# Retrieve original test set coordinates using stored indices
coordinates_test = data.loc[test_indices, ['Centroid', 'Centroid_y']]


# Create a DataFrame with actual and predicted values
results_df = pd.DataFrame({
    'X_Coordinate': coordinates_test['Centroid'],
    'Y_Coordinate': coordinates_test['Centroid_y'],
    'Actual_Greening_Potential': y_test.values,
    'Predicted_Greening_Potential': y_pred.flatten()
})


# Save the results DataFrame as CSV
results_df.to_csv('greening_potential_results_final_new.csv',
index=False)


# Trigger download (for Google Colab only)
files.download('greening_potential_results_final_new.csv')


# SHAP Analysis
explainer = shap.KernelExplainer(model.predict, X_train_scaled[:100])
# Using a subset of training data for the background
shap_values = explainer.shap_values(X_test_scaled[:100])  # Compute
SHAP values for a test set subset


# Reshape SHAP values if necessary
shap_values_correct = np.squeeze(np.array(shap_values))  # Ensure
correct dimensions
```

```
print(f"SHAP values reshaped: {shap_values_correct.shape}")  # Should match (100, number of features)


# Limit X_test_scaled to the first 100 rows to match SHAP values
X_test_scaled_subset = X_test_scaled[:100]  # Subset of test data to match SHAP values


# SHAP Summary Plot
plt.figure(figsize=(12, 6))
shap.summary_plot(shap_values_correct, X_test_scaled_subset, feature_names=X.columns, max_display=30)
plt.show()


# SHAP Dependence Plot for Within_building
plt.figure(figsize=(8, 6))
shap.dependence_plot('Within_building', shap_values_correct, X_test_scaled_subset, feature_names=X.columns)
plt.show()


# SHAP Waterfall Plot for a specific instance (e.g., the first test sample)
shap.plots._waterfall.waterfall_legacy(
    explainer.expected_value[0],
    shap_values_correct[0],
    X_test_scaled_subset[0],
    feature_names=X.columns
)
```

Decision Tree Model code used for Main Study

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.tree import DecisionTreeRegressor
from sklearn.preprocessing import StandardScaler
```

```
from sklearn.metrics import mean_squared_error, r2_score
import shap
import matplotlib.pyplot as plt
from google.colab import files  # Only if you're using Google Colab

# Assuming `data` is your DataFrame
features = [
    'OBJECTID', 'Centroid', 'Centroid_y', 'No_of_Trees', 'Tree_DIST', 'Crown_area',
        'Tree_h',   'NEAR_ANGLE',   'Within_building',   'Build_DIST', 'Built_up_impervious',
            'Street_DIST',     'Street_type',     'cycle_track_DIST', 'impervious_gridcode',
    'Urban_typology', 'NEAR_Green_space',
]

# Categorical features (One-Hot Encoding)
categorical_features = ['Public_area']
data_encoded = pd.get_dummies(data[categorical_features])

# Normalize NEAR_ANGLE to be within [0, 360]
data['NEAR_ANGLE'] = data['NEAR_ANGLE'] % 360

# Combine numerical and categorical features
X = pd.concat([data[features], data_encoded], axis=1)
y = data['Greening_potential']

# Split training and testing sets
X_train,   X_test,   y_train,   y_test   =   train_test_split(X,   y, test_size=0.2, random_state=42)
test_indices = X_test.index  # Store original test set indices

# Standardize the data
scaler = StandardScaler()
```

```
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Verify Within_building is in the dataset
assert 'Within_building' in X.columns, "Within_building is not in the feature list!"

# Define the Decision Tree Regressor
dt_model = DecisionTreeRegressor(max_depth=10, random_state=42)

# Train the model
dt_model.fit(X_train_scaled, y_train)

# Make predictions on the test set
y_pred = np.clip(dt_model.predict(X_test_scaled), 0, 100)

# Calculate Mean Squared Error and R^2 Score
mse = mean_squared_error(y_test, y_pred)
r2 = r2_score(y_test, y_pred)
print(f"Mean Squared Error: {mse}")
print(f"R^2 Score: {r2}")

# Retrieve original test set coordinates using stored indices
coordinates_test = data.loc[test_indices, ['Centroid', 'Centroid_y']]

# Create a DataFrame with actual and predicted values
results_df = pd.DataFrame({
    'X_Coordinate': coordinates_test['Centroid'],
    'Y_Coordinate': coordinates_test['Centroid_y'],
    'Actual_Greening_Potential': y_test.values,
    'Predicted_Greening_Potential': y_pred.flatten()
})
```

```
# Save the results DataFrame as CSV
results_df.to_csv('greening_potential_results_dt.csv', index=False)

# Trigger download (for Google Colab only)
files.download('greening_potential_results_dt.csv')

# SHAP Analysis
explainer = shap.Explainer(dt_model, X_train_scaled)  # Provide the model and scaled training data
shap_values = explainer(X_test_scaled)  # Compute SHAP values for the scaled test set

# SHAP Summary Plot
plt.figure(figsize=(12, 6))
shap.summary_plot(shap_values, X_test_scaled, feature_names=X.columns, max_display=30)
plt.show()

# SHAP Dependence Plot for Within_building
plt.figure(figsize=(8, 6))
shap.dependence_plot('Within_building', shap_values.values, X_test_scaled, feature_names=X.columns)
plt.show()

# SHAP Waterfall Plot for a specific instance (e.g., the first test sample)
shap.waterfall_plot(
    shap.Explanation(
        values=shap_values.values[0],
        base_values=shap_values.base_values[0],
        data=X_test_scaled[0],
        feature_names=X.columns
    ))
```

Appendix 2 | Structured Feature Engineered Dataset Used for Model Training

Table 12 | Description mapping of the Features used for model Training with the column names

Feature Description	Column Name
Unique identifier for each point feature.	OBJECTID
X-coordinates of the point.	Centroid_x
Y-coordinates of the point.	Centroid_y
Number of trees within the square grid.	No_of_Trees
Distance from the point to the nearest tree.	Tree_DIST
The crown area of the nearest tree.	Crown_area
Height of trees within the grid.	Tree_h
The angle of the location of the nearest tree with respect to the reference point.	NEAR_ANGLE
Field specifying whether the point is within a building.	Within_building
Distance from the point to the nearest building.	Build_DIST
The built-up imperviousness feature of the reference point.	Built_up_impervious
Distance from the point to the nearest street.	Street_DIST
The type of the nearest street to the point.	Street_type
The distance to the nearest cycle track.	cycle_track_DIST
The scale imperviousness of the location of the point of the study.	impervious_gridcode
Classification of Land Use Type / Urban Typology.	Urban_typology
The distance to the nearest green or open space.	NEAR_Green_space
Field specifying whether the location of the point of the study is public property.	Public_area
Greening potential of the grid (target variable for the model).	Greening_potential

OBJECTID	Centroid	Centroid_	No_of_Tree	Tree_DIST	Crown_ar	Tree_h	NEAR_ANGLE	Within_bu	Build_DIST	Built_up_i	Street_DIST	Street_typ	cycle_track	imperviou	Urban_typc	NEAR_Green_	Public_are	Greening_potential
1	694269	5334606	0	0	0	0	0	1	0.1	2	6.4946219	95	0	7	2.59	120.9112403	not_public	0
2	694279	5334606	0	0	0	0	0	1	0.1	2	1.894159	95	0	10	2.59	113.313441	not_public	0
3	694289	5334606	0	6.508403	0	0	40.39046153	1	0.1	2	2.7394773	95	0	7	2.59	105.7317442	not_public	16.02945475
4	694299	5334606	2	2.973278	49.06	14.19977	-140.762658	0	0	1	0.2179352	44	0	5	2.59	98.1500475	not_public	79.90140229
5	694309	5334606	2	3.743062	47.51	12.04352	-112.501528	0	2.573366	2	0	0	0	9	2.59	90.56835076	not_public	56.38954237
6	694139	5334616	0	0	0	0	0	1	0.1	2	0	0	0	10	3.32	235.5873354	Public_are	0
7	694149	5334616	0	0	0	0	0	1	0.1	2	0	0	0	10	3.32	226.7557994	Public_are	0
8	694159	5334616	0	0	0	0	0	0	0	1	0.9808656	44	1.9992235	10	3.32	218.0252109	Public_are	0
9	694169	5334616	0	0	0	0	0	0	0	1	0.5339612	94	0	10	3.32	209.4081961	Public_are	0
10	694179	5334616	0	0	0	0	0	0	0	1	1.7420953	99	0.7986776	10	3.32	200.9193684	Public_are	11.03232673
11	694189	5334616	1	2.4003	60.63	13.07379	-166.471493	0	0	2	0	0	0	7	3.32	192.5416977	Public_are	82.23595623
12	694199	5334616	0	0	0	0	0	1	0.1	2	0	0	0	10	4.9	184.1725975	not_public	5.952213247
13	694209	5334616	0	0	0	0	0	1	0.1	2	0	0	0	10	4.9	175.8034973	not_public	0
14	694219	5334616	0	0	0	0	0	1	0.1	1	0	0	0	10	4.9	167.4343972	not_public	6.335226063
15	694229	5334616	0	0	0	0	0	0	0	1	2.5173208	94	0	7	4.9	159.0656281	not_public	12.71411518
16	694239	5334616	0	6.573778	0	0	-17.5093833	0	0	1	6.4230699	94	0	3	4.9	150.8076922	not_public	71.32415793
17	694249	5334616	2	4.222625	63.45	16.10968	-152.070704	0	0	1	0	0	0	2	4.9	142.7727076	Public_are	100
18	694259	5334616	2	0.753715	101.64	15.87439	-119.530846	0	0	1	2.8969726	44	0	2	4.9	135.000489	Public_are	98.60727586
19	694269	5334616	1	5.41996	5.64	10.92633	63.42880293	0	0	1	1.7267996	44	0	2	2.59	127.4157109	not_public	26.6258326
20	694279	5334616	1	3.472805	6.75	4.389221	163.0871039	0	1.778971	2	6.3505718	44	0	4	2.59	119.8340141	not_public	6.840663915
21	694289	5334616	0	0	0	0	0	0	4.317166	2	1.2249309	44	0	5	2.59	112.2523174	not_public	4.108921004
22	694299	5334616	2	2.297471	50.21	12.57703	-34.7241751	0	0	2	4.0841447	94	0	5	2.59	104.6706207	not_public	74.23364482
23	694309	5334616	0	0	0	0	0	1	0.1	2	0	0	0	9	2.59	97.08892395	not_public	11.10520326
24	694009	5334626	0	0	0	0	0	1	0.1	2	0	0	0	10	3.81	186.5211508	not_public	0
25	694019	5334626	0	0	0	0	0	1	0.1	2	0	0	0	10	3.81	196.5141903	not_public	0
26	694029	5334626	0	0	0	0	0	0	0	1	2.8357687	44	0	10	3.81	206.5078134	not_public	0
27	694039	5334626	0	0	0	0	0	0	0	2	1.5222531	44	0	9	3.81	216.5020254	not_public	5.126450845
28	694049	5334626	0	5.223106	0	0	-89.4910004	0	0	2	2.2727625	44	0	9	3.81	226.4967483	not_public	22.3546811
29	694059	5334626	0	0	0	0	0	0	4.414021	2	1.0942762	44	0	9	3.81	236.4919174	not_public	0
30	694069	5334626	0	0	0	0	0	0	0	2	3.6230053	44	0	10	3.81	246.4874784	not_public	0

Figure 25 | Appendix - Dataset used for main study model training

OBJECTID	GRID_ID	Centroid	Centroid_Lat	Long	Street_FID	Near_street	Building_FID	Within_build	Near_Buildi	Land-Type	Tree_FID	NEAR_Tre	Tree_X	Tree_Y	Tree_Cou	Tree_are	Tree_h	Greening potential
1	WY-586	687569	5335096	48.14	11.521	5044	4.6943875	1165	0	1.9039474	mixed industrial or com	106	5.08411	687573	5335093	1	28.27	9.90405	0.1
2	WZ-586	687579	5335096	48.14	11.522	0	0	1165	1	0.1	mixed industrial or com	0	0	0	0	0	0	0	0.1
3	WM-585	687449	5335106	48.14	11.52	0	0	0	0	0	Network	80	2.92455	687450	5335109	2	87.47	18.2232	0.5
4	WN-585	687459	5335106	48.14	11.52	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1
5	WO-585	687469	5335106	48.14	11.52	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1
6	WP-585	687479	5335106	48.14	11.52	0	0	0	0	0	Network	79	6.15782	687481	5335112	0	0	0	0.1
7	WQ-585	687489	5335106	48.14	11.52	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1
8	WR-585	687499	5335106	48.14	11.521	0	0	0	0	0	open or green space	0	0	0	0	0	0	0	0.1
9	WS-585	687509	5335106	48.14	11.521	0	0	0	0	0	mixed industrial or com	89	3.22491	687510	5335109	1	83.16	18.1477	0.5
10	WT-585	687519	5335106	48.14	11.521	0	0	0	0	0	mixed industrial or com	90	4.41721	687515	5335109	1	193.83	20.0295	0.5
11	WU-585	687529	5335106	48.14	11.521	0	0	0	0	0	mixed industrial or com	102	0.86845	687528	5335106	1	212.22	15.4324	0.5
12	WV-585	687539	5335106	48.14	11.521	5044	4.3758581	0	0	0	mixed industrial or com	0	0	0	0	0	0	0	0.1
13	WW-585	687549	5335106	48.14	11.521	5044	1.5355914	0	0	0	mixed industrial or com	104	4.79383	687547	5335102	1	29.21	11.415	0.1
14	WX-585	687559	5335106	48.14	11.521	5044	2.2075302	0	0	0	mixed industrial or com	0	0	0	0	0	0	0	0.1
15	WY-585	687569	5335106	48.14	11.521	0	0	1165	0	2.8834056	mixed industrial or com	103	1.01321	687570	5335106	2	50.1	10.8576	0.1
16	WZ-585	687579	5335106	48.14	11.522	0	0	1165	1	0.1	mixed industrial or com	0	0	0	0	0	0	0	0.1
17	WA-584	687329	5335116	48.14	11.518	0	0	0	0	0	mixed industrial or com	0	0	0	0	0	0	0	0.1
18	WB-584	687339	5335116	48.14	11.518	0	0	0	0	0	mixed industrial or com	0	0	0	0	0	0	0	0.1
19	WC-584	687349	5335116	48.14	11.519	0	0	0	0	0	mixed industrial or com	0	0	0	0	0	0	0	0.1
20	WD-584	687359	5335116	48.14	11.519	1206	2.1041605	0	0	0	Network	0	0	0	0	0	0	0	0.1
21	WE-584	687369	5335116	48.14	11.519	3652	4.2382186	0	0	0	Network	0	0	0	0	0	0	0	0.1
22	WF-584	687379	5335116	48.14	11.519	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1
23	WG-584	687389	5335116	48.14	11.519	4315	4.515326	0	0	0	Network	0	0	0	0	0	0	0	0.1
24	WH-584	687399	5335116	48.14	11.519	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1
25	WI-584	687409	5335116	48.14	11.519	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1
26	WJ-584	687419	5335116	48.14	11.519	0	0	0	0	0	Network	0	0	0	0	0	0	0	0.1

Figure 26 | Appendix - Dataset used for pilot study model training approach 1

Appendix 3 | Street type data numerical encoding

Table 13 | Numerical encoding of street-type data

Street Type	Numerical encoding
none	0
track	21
bridleway	21
cycleway	43
trunk_link	43
footway	44
path	45
living_street	51
tertiary_link	53
steps	60
pedestrian	65
motorway	71
motorway_link	72
construction	76
secondary_link	77
unclassified	78
platform	86
residential	87
trunk	91
proposed	93
service	94
corridor	95
tertiary	95
primary_link	96
secondary	99
primary	99.5
bus_stop	99.8