Master Thesis

Big Data to Small Footprints: Predicting Office Operating Carbon Emissions Using Machine Learning

by

Stan Brouwer

(2671939)

First supervisor:	Name and surname
Daily supervisor:	Name and surname
Second reader:	Name and surname

July 10, 2025

Submitted in partial fulfillment of the requirements for the VU degree of Master of Science in Information Sciences

DECLARATION OF AUTHORSHIP

I, **Stan Brouwer**, declare that this thesis titled "Big Data to Small Footprint: Predicting Office Operating Carbon Emissions Using Machine Learning" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: Stan Brouwer

Date: 11 July 2025

Big Data to Small Footprints: Predicting Office Operating Carbon Emissions Using Machine Learning

Stan Brouwer
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
s.j2.brouwer@student.vu.nl

ABSTRACT

Your Master Thesis has to be written in English and should follow the ACM style (see the Overleaf template). We also suggest a *structured abstract* following the structure below.

Context. Write this at the end

Goal. Write this at the end

Method. Write this at the end

Results. Write this at the end

Conclusions. Write this at the end

As a very rough indication, the final thesis report typically entails, excluding appendixes, between 10 and 30 pages, with an average of 15 pages. The large variation depends on many variables including the specific field, project nature, and context. We advise to ask your supervisor if you should consider which number as a reference.

1 INTRODUCTION

Greenhouse gas (GHG) emissions associated with human activities have already caused 1.1 [0.95-1.20] °C of global warming above pre-industrial levels (IPCC, 2021), leading to irreversible changes to the climate system. In response, the Paris Agreement (2015) has established goals to limit global warming to well below 2°C and preferably 1.5°C. Yet, current policy outcomes fall short: under existing commitments, the world is on track for 2.8°C of warming by the end of the century (Emissions Gap Report, 2022; IPCC, 2024). To avoid this outcome, great reductions in emissions are required, with the UNEP (2024) noting that achieving the 1.5°C target requires decreasing current emissions by 42

Within this context, corporations face growing pressures to align their operations with climate goals. Governments are introducing stricter regulations such as the EU's Corporate Sustainability Reporting Directive (CSRD, 2024), which mandates larger companies to disclose climate-related risks, GHG emissions and sustainability. At the same time, societal expectations are shifting: public awareness and concern about climate change creates reputational incentives for companies to demonstrate environmental responsibility (Zheng et al., 2024, Flammer, 2013). Investors too are increasingly valuing environmental, social and governance (ESG) factors into their decision-making, although the extend is unclear(x), favouring organizations with sustainability strategies (Harasheh, Bouteska Manita, 2024; Famiyeh Kwarteng, 2017, Shukla et al., 2009; Cheng et al., 2014). Finally, legal accountability is on the rise. Landmark rulings such as Milieudefensie v. Royal Dutch Shell underscores the growing role of the judiciary in enforcing corporate climate responsibility, further compelling companies to act.

The role of office buildings. Office buildings present a significant opportunity for GHG emissions reduction via improved energy management (UNEP, 2024). Buildings account for approximately 40

Teng Ying (2023) estimate that energy use in building operations alone contributes 19

Managing and reducing energy consumption starts with understanding it. This begins with quantifying energy usage, which is typically normalized for floor area and period and expressed as energy use intensity (EUI, kWh/m2/year). (Hong et al. 2015; Nikolaou et al. 2015). Generally, there are two methodological approaches to assess energy performance: physics-based simulations which use detailed structural and environmental data in combination with energy simulation tools (EnergyPlus, TRNSYS, ESP-r) to calculate consumption; or data driven / statistical models, which infer patterns from historical building and environmental data to predict energy use (Seyedzadeh et al., 2018 (also other paper)

While simulation-based models are widely used in energy optimization and certification schemes, they are computationally intensive and often impractical for large-scale or early-stage evaluations. Optimization algorithms that rely on simulation tools suffer from high time overhead, making them costly to scale or apply across portfolios (Ascione et al., 2014; Smarra et al., 2018). In contrast, data-driven models offer a more scalable alternative, reducing costs and enabling rapid evaluation of energy performance across large datasets (Dounis Caraiscos, 2009; Deb et al., 2016). They are also increasingly used to support design decisions, optimize HVAC systems, and benchmark energy performance (Zhao Magoulès, 2012b; Ahmad et al., 2014). Importantly, classification-based techniques can facilitate analysis of complex variables like occupant behavior or usage patterns, which are difficult to model through simulation (Yu et al., 2011).

Practical barriers. While technological solutions such as smart meters, advanced HVAC systems, and retrofitting exist and are often cost-effective, their implementation remains limited (Astmarsson et al., 2013). Thus, the challenges are informational and economic rather than technical. Reliable performance data is often missing or non-transparent, preventing markets from functioning efficiently (Hsu, 2014). This hinders investors and tenants from valuing energy efficiency, reducing incentives to invest in sustainable buildings.

Transparency initiatives such as the EU's Building Energy Performance Directive (EPBD), ENERGY STAR, and LEED have improved information availability. Studies show certified buildings command rental and sales premiums (Eichholtz et al., 2009; Fuerst McAllister, 2011; Miller et al., 2008). Yet these certifications are inconsistent across regions, not always publicly accessible, and often rely on complex, expert-driven simulations.

Practical constraints further hinder sustainability initiatives. Office buildings are often multi-tenant and rely on shared energy meters. Tenants are often charged based on floor area rather than actual usage. This structure limits the feedback tenants receive on their actual energy use and undermines incentives for reducing consumption (Kühn et al., 2024). At portfolio level, building owners and real estate managers struggle to collect and compare energy performance data, making it difficult to identify underperforming assets, comply with ESG mandates, or justify sustainability investments. Providing performance data has broader economic benefits as well as it reduces adverse selection, aligns stakeholder incentives, and enables more targeted, cost-effective policies (Akerlof, 1970; McKinsey, 2009).

Motivation. With the increased attention to decarbonization, there is a growing body of research on energy consumption in buildings. Much of it focuses on the residential sector, while office buildings remain relatively under-studied. As office buildings have distinct characteristics that differentiate them from residential structures , they require separate analysis.

This research contributes to the scientific literature by applying ML/statistical techniques to large datasets of over 7000 US office buildings. We evaluate different machine learning and statistical methods that estimate the energy consumption of US office buildings. Novel is the combination with weather data based on location and size of the considered dataset.

The study is structured around the following research questions: RQ1: What are the key characteristics (e.g. building size, age, number of occupants, operating hours) most strongly determining energy use in US office buildings? RQ2: How do different modelling techniques compare in accuracy when used to predict yearly energy consumption?

By developing and assessing the accuracy of predictive models, this study aims to offer a faster and more scalable approach to support effective energy management, targeted retrofits, and strategic decision-making in the commercial real estate sector.

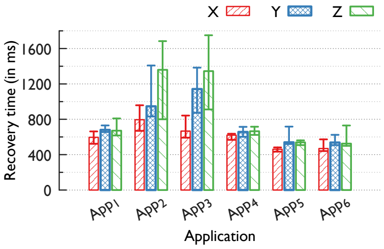


Figure 1: ***Simple one-column figure. Please include a brief explanation or takeaway.***

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

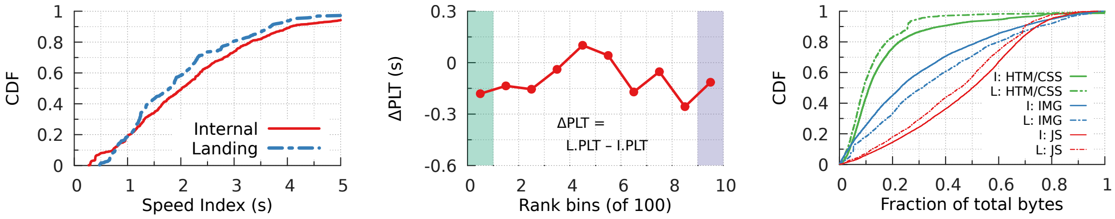
Table 1: ***A simple table describing the characteristics of a data set or the results of an experiment***

Char.	***#samples***	***Count***	***Perf. Score***		
		of items	***X***	***Y***	***Z***
P	214	56	9	23	24
Q	117	27	7	10	10
R	222	11	6	4	1
S	187	9	1	6	2
T	180	16	7	5	4

2 BACKGROUND

The energy consumption of buildings, and in particular office buildings, contribute significantly to the overall energy usage and CO2 emissions (Teng Yin, 2023; Wu et al., 2011). Consequently, developing accurate energy consumption prediction models has become crucial for optimizing energy usage, reducing costs, and promoting sustainability in the office environment (Mariano-Hernández et al., 2020). While scholars have recognized the importance of energy prediction for achieving energy efficiency and cost savings since the mid-1980s (Chammas et al., 2019), this literature review aims to analyze the current state of research on energy consumption in highly energy efficient office buildings, emphasizing the key drivers and assess the integration of sensor data in predictive modeling methods

Drivers of office building energy usage. Energy consumption in buildings is influenced by many physical, operational, behavioural and environmental variables. Physicists and engineers tend to create physical models calculating heat dissipation and radiation, while the data-driven domains use big-data approaches.



(a) **(b)** **(c)**

Figure 2: ***Generate clear and beautiful figures (in PDF) that can be rendered side by side while still being easy to read and interpret. Choose colors wisely from the colorbrewer2.org website.***

From an architectural perspective, studies have emphasized the importance of building envelope properties (insulation, wall-to-window ratio. air-tightness), HVAC system efficiency, and the impact of passive design features (orientation, shading, thermal mass) on energy demand (Perez-lombar et al., 2008; dodoo et al., 2010).

In the behavioral sciences, researchers have identified occupant behaviour and organizational culture as significant drivers of energy use variations. Gram-Hanssen (2010) and Santin and colleagues (2009) show that user habits such as thermostat settings, window opening and equipment usage can lead to large differences in energy performance, even among technically equivalent buildings. This "performance gap" is often attributed to behavioral unpredictability.

Empirical studies have attempted to quantify the relative impact of different energy consumption factors, often using regression models clustering or machine learning. Most researched variables are discussed below:

Floor area. Nearly all analyses find larger offices use more energy, roughly in proportion. A Korean nationwide study reports that floor area alone explains approximately 90% of the variation in energy usage of offices (R2 = 0.893 - 0.909; Kim Kim, 2020). Often, regression models include floor area. Based on a New York city office dataset, Kontokosta reports that each additional 1,000 ft2 of area raised source energy usage by about 0.064 kBtu/ft2 ($p < 0.05$; 2012). Although the energy usage scales roughly linearly with floor area, it is hypothesized that larger office buildings benefit from economies of scale and have lower energy consumption per floor area (energy intensity). Kim Kim (2020) found that the coefficient of determination between the total floor area and energy consumption was higher with a quadratic model for large office buildings than with a linear model, indicating possible scale benefits. However, the authors also note that the linear regression model performed better for smaller offices and suggest the application of segmented regression. These interactions could be further complicated by the interactions between building age, energy rating (newer buildings tend to be better insulated), and building age and size (newer buildings tend to be larger), but the dominant pattern is clear: larger buildings use more energy.

Number of occupants. Studies consistently find that energy intensity rises with how many people are inside of a building. For instance by including workers per area as a positive factor (Sharp, 1996). In the paper discussing the New York city data, Kontokosta found that adding one worker per 1.000 ft2 increased the annual energy usage with about 10.48 kBtu/ft2 ($p > 0.01$).

Operating hours. The longer an office is used per week, the more energy it consumes. Regression studies often include operating hours as a predictor. For example, Sharp (1996) identified operating hours as a strong driver for energy usage. From the New York city dataset, Sharp found that each extra open hour per week corresponded with an energy usage increase of 0.447 kBtu/ft^2 ($p < 0.05$). A technical report by the ENERGY START RATING TEAM (2019) reports 0.614 kBtu/ft^2 ($p <0.0001$).

Building age. Many regressions find that newer offices do not use less area on a per-area basis. The New York city study (reference) found a negative correlation of energy usage with age: offices over 80 years of age used approximately 30% less energy per ft^2 than the average office (Kontokosta, 2012). This trend is also observed in a UK study, reporting electricity intensity is higher in recently-built offices (Godoy-Shimizu et al., 2018). The Building Energy Research Centre of Tsinghua University (2023) attributes the observed trend of rising energy intensity of office buildings mainly to the rise of air conditioning systems.

Weather influences. Weather significantly affects commercial building energy use, which has to be accounted for. The simplest approach is by considering the degree-days: summing the differences between outdoor temperature and a base temperature, loosely corresponding to the temperature gradient between the indoor and outdoor. (Makhmalbaf, Srivastava Wang, 2013) More refined methods fit regression models of energy versus weather variables. With the rise of machine learning more advanced methods such as neural networks, random forests etc. can model nonlinear responses to temperature and other factors. Each approach to consider the influence of weather uses common meteorological features such as temperature, humidity, solar radiation and wind speed, to estimate what the energy use would have been under a reference climate. By removing weather effects, these models enable better year-over-year or between-building comparisons (Liu et al., 2023).

The most common normalization method to adjust for the effects of weather is the heating and cooling degree-day (HDD, CDD) method (Akander, Alvarez Johannesson, 2004), although the methodology to normalize for other variables is similar. Degree-days represent the total positive or negative differences between a set temperature and the average temperature for a given period of time

(ASHRAE, 2009), should I include formula? DD =(To - Tb) which has been specified as 18,3 °C or 65 °F in the U.S. Kissock et al. (2004) discuss the variable-base degree-day (VBDD) method, in which the most optimal base temperature which provides the best statistical fit is calculated. In the basic degree-day method, a regression model correlates energy use with degree days, which is then evaluated on its R value. For the VBDD, multiple models with different base temperatures are developed, and the model with the highest R value is selected.

3 METHODS

This study develops predictive models for the annual energy use intensity (EUI) of office buildings by using building characteristics and climate data. The methodology is structured into four main sections: are organized into four main sections: Data Collection Preprocessing, Feature Selection, Model Development, and Model Evaluation. An a priori significance level of 0.05 was applied throughout.

3.1 Data collection preprocessing

Data were compiled from 26 publicly available datasets containing annual energy consumption and associated building characteristics [citation]. Only records explicitly labeled as office buildings were retained. Further filtering ensured inclusion only of entries with non-missing, positive values for the following variables: site energy use (defined as total annual site energy consumption in MWh), floor area, year built, energy source indicators, and location (city or state).

All imperial units were converted to SI units except for time and energy use (reported in MWh for interpretability). Energy intensity (EUI) was calculated as site energy use divided by floor area. Variables and their units are detailed in Table X.

Climate data (Heating Degree Days [HDD] and Cooling Degree Days [CDD]) were sourced from the National Weather Service Climate Prediction Center [citation], which provides monthly aggregated degree day statistics for 359 major US cities. Building locations were matched to city-level climate data; if city data were unavailable, the state-level population-weighted average was applied. The final dataset includes 32,686 yearly observations across 7,226 unique office buildings, spanning 2010 to 2023.

3.2 Feature selection

Predictor variables were initially assessed for compliance with the assumptions of normality, homoscedasticity, and independence. Distributions were visually inspected via histograms and Q-Q plots. Skewness and kurtosis z-values were calculated, with ±2 as the threshold for normality. Shapiro-Wilk tests were performed on random samples (n=2,000 per variable) to mitigate sensitivity to large sample sizes (Field, 2009). For non-normal variables, log-transformations were applied and reassessed. Homoscedasticity was evaluated with Levene's test.

3.3 Correlation analysis

The association between candidate predictors and yearly energy use intensity (EUI) was examined using Pearson's correlation coefficient(PCC), or Spearman's rho in case of non-normality. Cohen's (1988) guidlines on the interpretation of correlation coefficients served as reference, where correlation values of r = 0,1, 0,3 and 0,5 indicate small, moderate or high correlation. However, it should be repeated that the interpretation of the coefficients ultimately depends on their context and purpose [source], and given the relatively large sample size in our dataset (n > 30.000), even relatively minor correlation coefficients can achieve statistical significance and reflect real associations.

3.4 Multicollinearity assessment

An inter-variable correlation matrix and Variance Inflation Factors (VIF) were calculated to detect multicollinearity. Variables with VIF > 5 were considered for removal or combination to ensure model stability.

3.5 Development of predictive models

3.5.1 Multiple linear regression (OLS). A simple linear parametric model was developed to predict the energy use of electric-only buildings. A separate multivariate linear regression model was developed to predict both the electric as fuel energy use for mixed-source buildings. Assumptions of normality, homoscedasticity, and independence were assessed via visual inspection of the histogram and Q-Q plots, and with Shapiro-Wilk's and Levene's tests. The general form was: EUIi=0+1FloorAreai+2YearBuilti+3OperatingHoursi+4CDDi+5HDDi+i

3.5.2 Decision tree regression. A decision tree model was developed to capture potentially non-linear relationships between building characteristics and energy use. Unlike linear models, decision trees recursively split the data into smaller groups based on decision 'rules' that maximize the difference in the outcome variable between branches. This allows the model to 'learn' threshold effects and interactions that might be difficult to represent parametrically. To tune the model and prevent overfitting, key hyperparameters were optimized using 5-fold cross-validation. such as the maximum number of splits (tree depth), the minimum number of samples required to split a node (...) were optimized using 5-fold cross validation. In this approach, the training data is split into five equal subsets (folds). The model is trained on four sets, and validated on the remaining set, the process being repeated 5 times with rotation in order for each set to be used as a validation set exactly once. The average performance across the iterations provide a robust estimate of how well the model generalizes to new data.

Before hyperparameter tuning, we randomly split the dataset into 80% training data and 20% testing data (holdout-set), which was not used for the 5-fold cross validation. We tested each combination of hyperparameters with tree depths ranging from 2 to 20 and minimum samples per split ranging from 2 to 20. Model accuracy was evaluated using mean absolute error (MAE), which measures the average absolute difference between predicted and actual annual energy values. The hyperparameter combination with the lowest average MAE across validation folds was selected. The final decision tree was then retrained on the full training dataset using this optimal configuration and tested on the unseen holdout set.

3.5.3 Random forest regression. To improve prediction accuracy and mitigate overfitting observed in single decision trees, a random forest model was implemented. This ensemble method builds multiple decision trees using bootstrapped samples and random feature selection, averaging their predictions to enhance generalization. Hyperparameter tuning followed the same 5-fold cross-validation and holdout approach. Feature importance metrics were extracted to interpret predictor influence

3.6 Model evaluation

Linear regression models were evaluated using Rsquared, adjusted Rsquared MSE, RMSE, and MAE on the full dataset. Nonlinear models were assessed on the 20

Evaluation metrics include:

Root Mean Squared Error (RMSE): Square root of average squared differences between predicted and actual values, penalizing larger errors [46].

Mean Absolute Error (MAE): Average absolute difference between predictions and observations, providing a unit-consistent error measure.

Mean Absolute Percentage Error (MAPE): Average percentage deviation between predictions and actual values, enabling scale-independent comparison [45].

Mean Bias Error (MBE): Average bias indicating systematic over- or underprediction.

Coefficient of Variation of RMSE (CV$_RMSE) : RMSEnormalizedbythemeanofobservedvalues, allowingcomparabilityacrossdatasets$[14].

Coefficient of Determination (Rsquared)): Proportion of variance in the dependent variable explained by the model.

Where: At = actual value P2 = predicted value n = number of observations A = mean of actual values

4 RESULTS

4.1 Statistics

The considered dataset contained 32.686 valid yearly observations across 7.226 unique office buildings, with observation dates ranging from 2010 to 2023. Descriptive statistics are reported in table X.

Results of the normality tests are presented in table X. In combination with the visual inspection of the histograms and Q-Q plots it was assumed that none of the variables demonstrated normality.

4.2 Feature selection

4.3 Linear regression model

4.4 Model comparison

5 DISCUSSION

Here you put your results in context (possibly grouped by research question). Usually, this section focuses on analyzing the implications of the proposed work for current and future research and for practitioners.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris

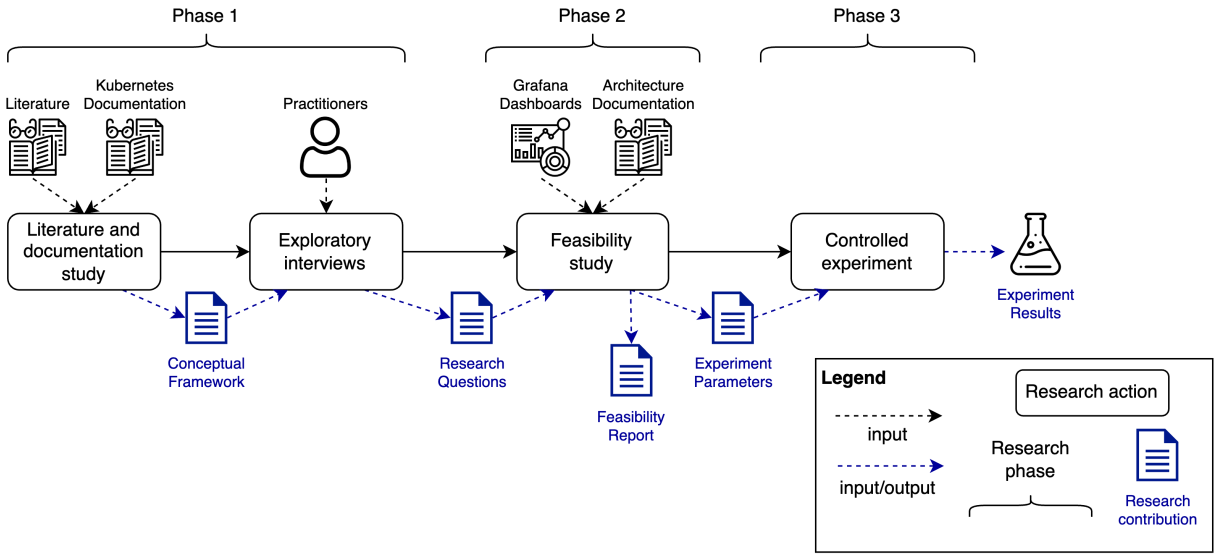


Figure 3: ***Overview: Research strategy, research methods and results. [Credits: illustration with Diagram.net]***

felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Donec odio elit, dictum in, hendrerit sit amet, egestas sed, leo. Praesent feugiat sapien aliquet odio. Integer vitae justo. Aliquam vestibulum fringilla lorem. Sed neque lectus, consectetuer at, consectetuer sed, eleifend ac, lectus. Nulla facilisi. Pellentesque eget lectus. Proin eu metus. Sed porttitor. In hac habitasse platea dictumst. Suspendisse eu lectus. Ut mi mi, lacinia sit amet, placerat et, mollis vitae, dui. Sed ante tellus, tristique ut, iaculis eu, malesuada ac, dui. Mauris nibh leo, facilisis non, adipiscing quis, ultrices a, dui.

6 LIMITATIONS (OR THREATS TO VALIDITY)

Report about each type of limitations of your study, or threat to the validity of aspects of its design or execution, and how did you mitigate them, according to the classification framework that fits best your study design. For instance, for empirical experiments, use the one proposed by Wohlin *et al.* [2]. Accordingly, the threats as organized as the following sections. You may also use the work of Verdecchia *et al.* [1] to present this section at best, for example, if discussing the limitations is a better fit.

6.1 Internal Validity

6.2 External Validity

6.3 Construct Validity

6.4 Conclusion Validity

7 RELATED WORK

Describe here scientific papers similar to your thesis work, both in terms of goal and methodology. One paragraph for each paper (we expect about 5-8 papers to be discussed). Each paragraph contains: (i) a brief description of the related paper and (ii) a black-on-white description about how your work differs from, or overlaps with, the related paper, hence emphasizing the novelty contributed by this thesis. You may place this section immediately after the Background section, if necessary.

Sed lobortis, justo et pretium lobortis, mauris turpis condimentum augue, nec ultricies nibh arcu pretium enim. Nunc purus neque, placerat id, imperdiet sed, pellentesque nec, nisl. Vestibulum imperdiet neque non sem accumsan laoreet. In hac habitasse platea dictumst. Etiam condimentum facilisis libero.

8 CONCLUSION

Briefly summarize your contributions, and share a glimpse of the implications of this work for future research.

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

REFERENCES

[1] Roberto Verdecchia, Emelie Engström, Patricia Lago, Per Runeson, and Qunying Song. 2023. Threats to validity in software engineering research: A critical reflection. *Information and Software Technology* 164 (Dec. 2023), 107329.

[2] C. Wohlin, P. Runeson, M. Höst, M.C. Ohlsson, B. Regnell, and A. Wesslén. 2012. *Experimentation in Software Engineering - An Introduction.* Kluwer Academic Publishers, International Series in Software Engineering.