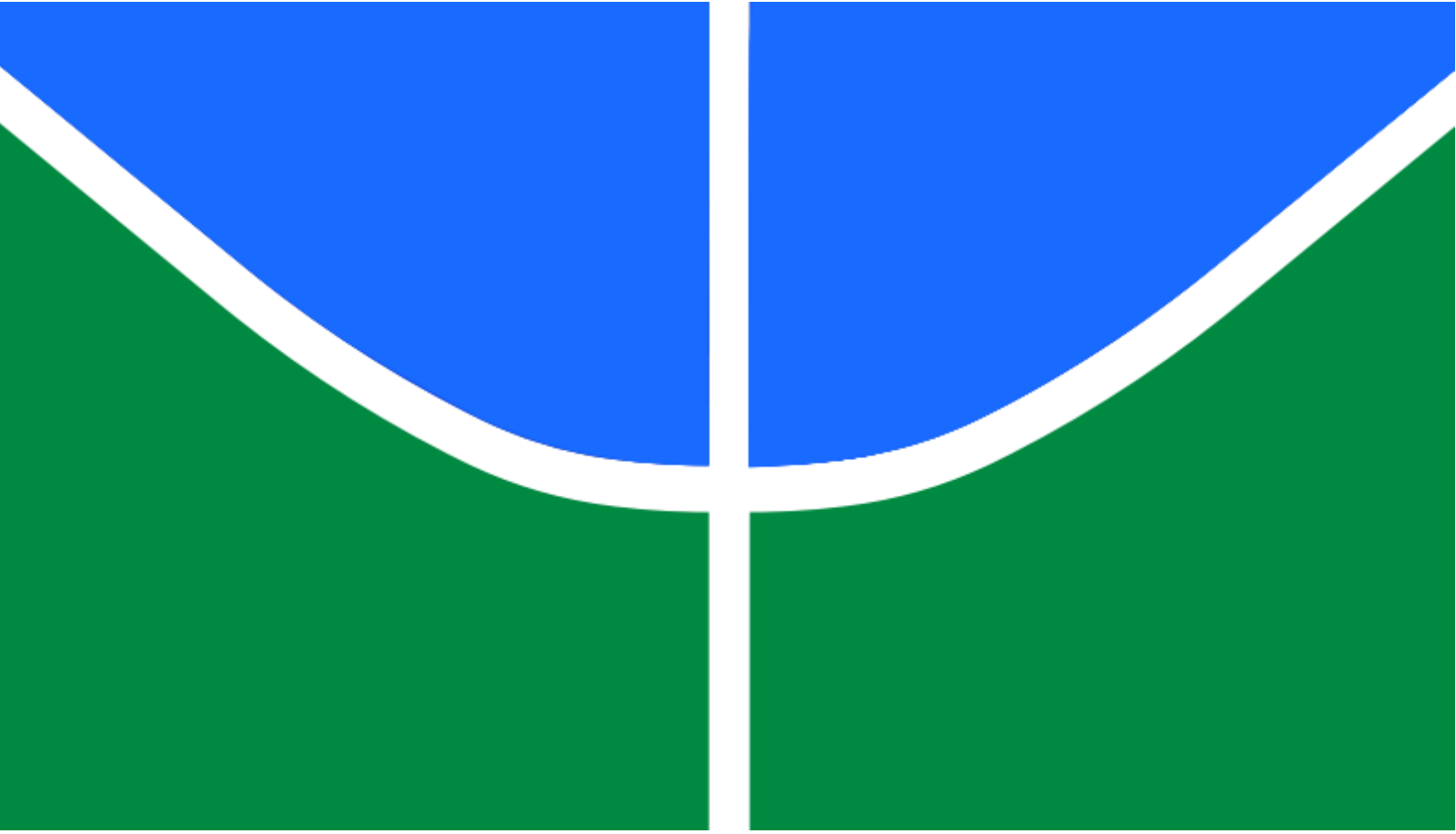


ARTIFICIAL NEURAL NETWORKS WITH BAYESIAN OPTIMIZATION AND FEATURE SELECTION APPLIED TO REGRESSION IN ENGINEERING PROBLEMS

RUAN DE ALENCAR CARVALHO

DISSERTAÇÃO DE MESTRADO EM ESTRUTURAS E CONSTRUÇÃO CIVIL
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

FACULDADE DE TECNOLOGIA

UNIVERSIDADE DE BRASÍLIA

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA

DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

ARTIFICIAL NEURAL NETWORKS WITH BAYESIAN OPTIMIZATION AND FEATURE SELECTION APPLIED TO REGRESSION IN ENGINEERING PROBLEMS

RUAN DE ALENCAR CARVALHO

ORIENTADOR: FRANCISCO EVANGELISTA JÚNIOR, PhD.

DISSERTAÇÃO DE MESTRADO EM ESTRUTURAS E CONSTRUÇÃO CIVIL

BRASÍLIA/DF: MARÇO/2025

UNIVERSIDADE DE BRASÍLIA

FACULDADE DE TECNOLOGIA
DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL

ARTIFICIAL NEURAL NETWORKS WITH BAYESIAN OPTIMIZATION AND FEATURE SELECTION APPLIED TO REGRESSION IN ENGINEERING PROBLEMS

RUAN DE ALENCAR CARVALHO

DISSERTAÇÃO SUBMETIDA AO DEPARTAMENTO DE ENGENHARIA CIVIL E AMBIENTAL DA UNIVERSIDADE DE BRASÍLIA COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM ESTRUTURAS E CONSTRUÇÃO CIVIL.

APROVADA POR:

_____

FRANCISCO EVANGELISTA JÚNIOR, *PhD.* (ENC – UnB)
(ORIENTADOR)

_____

LUCIANO MENDES BEZERRA, PhD. (ENC-UnB)
(EXAMINADOR INTERNO)

_____

AMÉRICO BARBOSA DA CUNHA JÚNIOR, DSc. (LNCC)
(EXAMINADOR EXTERNO)

_____

RAMON SALENO YURE RUBIM COSTA SILVA, DSc. (ENC-UnB)
(SUPLENTE)

BRASÍLIA/DF, 17 DE MARÇO DE 2025.

FICHA CATALOGRÁFICA

CARVALHO, RUAN DE ALENCAR
ARTIFICIAL NEURAL NETWORKS WITH BAYESIAN OPTIMIZATION AND FEATURE SELECTION APPLIED TO REGRESSION IN ENGINEERING PROBLEMS [Distrito Federal] 2025.

xx, 127 p., 210 x 297 mm (ENC/FT/UnB, Mestre, Estruturas e Construção Civil, 2025).
Dissertação de Mestrado - Universidade de Brasília. Faculdade de Tecnologia.
Departamento de Engenharia Civil e Ambiental.

1. Pavimento de concreto	2. Conexões parafusadas
3. Aço formado a frio	4. Aprendizado de máquina
5. Redes neurais artificiais	
I. ENC/FT/UnB	II. Título (Mestre)

REFERÊNCIA BIBLIOGRÁFICA

CARVALHO, R.A. (2025). Artificial neural networks with Bayesian optimization and feature selection applied to regression in engineering problems. Dissertação de Mestrado em Estruturas e Construção Civil, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, DF, 147 p.

CESSÃO DE DIREITOS

AUTOR: Ruan de Alencar Carvalho
TÍTULO: Artificial Neural Networks with Bayesian Optimization and Feature Selection applied to Regression in Engineering Problems
GRAU/ANO: Mestre/2025

_____

Ruan de Alencar Carvalho
runcarvalho@gmail.com
Brasília/DF - Brasil

AGRADECIMENTOS

Agradeço aos meus pais, Osvaldo e Nerede, por sempre acreditarem no meu potencial, me incentivarem a sair da zona de conforto e me apoiarem em minhas decisões. Jamais teria chegado onde cheguei sem o suporte dos dois.

À minha irmã, Alícia, por dividir a casa comigo, trocando conversas e me fazendo companhia diariamente. Aos amigos de sempre por me ajudarem a esquecer dos estresses do Mestrado.

Especialmente, à minha namorada, Maria Vitória, por ser compreensiva e parceira em todos os momentos que me ausentei do relacionamento para tratar de demandas do Mestrado. Me acompanhou em todos os 2 anos de curso e tornou toda a jornada mais leve.

Ao meu orientador, Francisco Evangelista, pelas grandes oportunidades a mim proporcionadas, além de muita confiança e paciência no processo de orientação profissional.

À Universidade de Brasília (UnB) e ao governo brasileiro por fornecerem educação de nível superior de qualidade e gratuita. À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e à Fundação de Apoio à Pesquisa do Distrito Federal (FAP-DF) pelo suporte financeiro e pelo apoio à pesquisa e desenvolvimento tecnológico.

Ao Laboratório de Métodos Computacionais e Inteligência Artificial (LAMCIA) pela infraestrutura necessária para desenvolvimento da pesquisa.

E, por fim, a mim mesmo, por ter sido resiliente nesse curto e intenso período que é o Mestrado. Foram 2 anos de muito aprendizado e crescimento pessoal. Sou muito grato ao eu do passado por não ter desistido e ter acreditado que grandes resultados requerem grandes esforços.

ABSTRACT

The present work consists of an implementation of Data-Driven Modeling (DDM) with Machine Learning (ML) techniques, in particular Artificial Neural Networks, as an approach to the study of net section capacity of cold-formed Steel Bolted Angles (SBA) under axial tension; concrete pavement surface temperature; and difference in temperature from surface to bottom of pavement concrete slabs. Two other datasets were collected from a known public repository to compare other authors' results with the ones obtained in the present thesis. The datasets are composed of instances from experimental tests, numerical simulations, and analytical and standards' equations. The present thesis aims to obtain accurate models capable of surrogating these methods with reduced financial and computational costs.

A baseline methodology was presented with 5-fold cross-validation, hyperparameter tuning with Bayesian Optimization (BO) and regularization techniques, such as early stopping, weight decay and batch normalization. Different sets of input variables were used, implementing feature selection and feature importance algorithms and aiming to achieve sufficiently accurate models with a reduced number of inputs. Besides that, another methodology was used, a BO-based algorithm that was named SOFAH, Simultaneous Optimization of Feature Augmentation and Hyperparameters. It addresses simultaneously and automatically the optimization of hyperparameters and feature selection of an augmented set of the original features. SOFAH aims to further improve the baseline models while providing information on important features for predictions of the response variable.

The developed models, both on gathered datasets and on public repository data, achieved very high accuracy and provided information on most relevant features for each analyzed engineering problem. The overall results of the present work show the effectiveness of ML and DDM as powerful tools for studying and predicting physical properties relevant to engineering problems.

Keywords: machine learning; artificial neural networks; Bayesian optimization; climate data; concrete pavements; steel connections.

RESUMO

REDES NEURAIS ARTIFICIAIS COM OTIMIZAÇÃO BAYESIANA E SELEÇÃO DE VARIÁVEIS APLICADAS NA REGRESSÃO DE PROBLEMAS DE ENGENHARIA

O presente trabalho consiste na implementação de Modelagem Baseada em Dados (MBD) com técnicas de Aprendizado de Máquina (AM), especificamente Redes Neurais Artificiais (RNA), como uma abordagem ao estudo da resistência da seção líquida de cantoneiras de aço parafusadas sob tração axial; da temperatura da superfície de pavimentos de concreto; e da diferença de temperatura entre a superfície e a base de lajes de concreto de pavimentos. Dois outros conjuntos de dados foram coletados de um repositório público conhecido para comparar os resultados de outros autores com os obtidos na presente dissertação. Os conjuntos de dados são compostos por amostras de testes experimentais, simulações numéricas e equações.

Uma metodologia de base foi apresentada com validação cruzada 5-fold, ajuste de hiperparâmetros com Otimização Bayesiana (OB) e técnicas de regularização, como *early stopping*, decaimento de peso e *batch normalization*. Diferentes conjuntos de *inputs* foram utilizados, implementando algoritmos de seleção de atributos e importância de atributos, com o objetivo de alcançar modelos de acurácia satisfatória com um número reduzido de entradas. Além disso, outra metodologia foi utilizada, um algoritmo baseado em BO denominado SOFAH, *Simultaneous Optimization of Feature Augmentation and Hyperparameters*. Ele realiza simultaneamente a otimização de hiperparâmetros e a seleção de *inputs* das RNAs. O SOFAH visa aprimorar os modelos da metodologia de base, ao mesmo tempo que fornece informações sobre as variáveis importantes para a previsão da variável de resposta.

Os modelos desenvolvidos, tanto nos conjuntos de dados coletados quanto em dados de repositórios públicos, atingiram alta acurácia e proveram informações acerca das variáveis mais relevantes. Em geral, os resultados deste trabalho demonstram a eficácia das técnicas de AM e da modelagem baseada em dados como ferramentas ponderosas para o estudo e previsão de propriedades físicas relevantes para problemas de engenharia.

Palavras-chave: pavimento de concreto; conexões parafusadas; aço formado a frio; aprendizado de máquina; redes neurais artificiais.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND SYMBOLS

Acronyms

AISI	American Iron and Steel Institute
ANN	Artificial Neural Network
APE	Absolute Percentage Error
BO	Bayesian Optimization
BSA	Bolted Steel Angle
CEN	European Committee for Standardization
CL	Cooling Load
CV	Coefficient of Variation
DDM	Data-Driven Modeling
ENEFF	Energy Efficiency dataset
FEM	Finite Element Method
FS	Feature Selection
GP	Gaussian Process
HL	Heating Load
HPO	Hyperparameter Optimization
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLP	Multi-Layer Perceptron
MSE	Mean Squared Root
PFI	Permutation Feature Importance
ReLU	Rectified Linear Unit
RMSE	Root Mean Squared Error

SD	Standard Deviation
SOFAH	Simultaneous Optimization of Feature Augmentation and Hyperparameters
WM	Wang-Mendel
YAH	Yacht Hydrodynamics dataset

Symbols

d_h	Bolt hole diameter
d	Nominal bolt diameter
P	Applied load
P_{nt}	Applied load in net section failure
b_c	Width of connected leg
b_d	Width of unconnected leg
t	Thickness
n_{lines}	Number of bolt lines
n_{holes}	Number of holes per line
$\bar{x}$	Connection eccentricity in x axis
$\bar{y}$	Connection eccentricity in y axis
L	Length of connection in z axis
L_t	Length of connection in y axis
F_u	Ultimate strength of the material
T_{obs}	Observed resistance of net section
A_{nt}	Net section cross-sectional area
x_g	Horizontal coordinate of centroid of cross-section
A_i	Area of i-th element composing cross-section decomposition
e_2	Distance from center of bolt hole to edge in z axis

p_1	Distance of consecutive bolts in z axis
β_2	Reduction factor for 2-bolted steel angles
β_3	Reduction factor for 3-bolted steel angles
n_b	Number of bolts in the cross-section
D	Current day
H	Current hour
RH	Relative humidity
ff	Wind speed
T_{surf}	Pavement surface temperature
T_{surf}^{pred}	Predicted pavement surface temperature
T_{air}	Air temperature
ΔT	Difference in temperature from surface to bottom of concrete slab
h_1	Thickness of pavement concrete slab
h_2	Thickness of pavement base
q_{net}	Net rate of heat flow
q_s	Energy absorbed from direct solar radiation
q_a	Energy absorbed from diffuse radiation
q_c	Energy transferred due to convection
q_k	Energy transferred due to conduction
q_r	Energy emitted through outgoing radiation
α_a	Surface absorptivity
R_0	Solar constant
τ_a	Transmission coefficient
z	Zenith angle
i	Angle between normal to the surface and direction of radiation

ϕ_l	Latitude
δ_s	Solar declination
h	Hour angle
σ_0	Stefan-Boltzman constant
ρ	Vapor pressure
k	Thermal conductivity
h_d	Depth on concrete slab
T_d	Temperature at depth h
T_m	Average of surface and air temperatures
U	Average daily wind velocity
ε	Emissivity
h_c	Surface coefficient of heat transfer
θ	Model parameters
$\hat{\theta}$	Estimates of model parameters
$J(\theta)$	Loss function
$\nabla_\theta J(\theta)$	Gradient of loss function with respect to parameters
α	Learning rate
$\hat{y}$	Predictions
y	Target variable
x	Inputs
k	Number of folds in k-fold cross-validation
f_a	Activation function
w	Neural network weight
b	Neural network bias
n_l	Number of inputs of layer l

γ_1	Exponential decay rate for first moment estimate
γ_2	Exponential decay rate for second moment estimate
m	Biased first moment estimate
v	Biased second moment estimate
$\widehat{m}$	Bias-corrected first moment estimate
v	Bias-corrected second moment estimate
δ	Numerical stability constant
γ	Scaling parameter in batch normalization
β	Shifting parameter in batch normalization
y_N	Training set targets
y_{N+1}	Unseen target
β_G	Precision of Gaussian noise
δ_{nm}	Kronecker delta
$m(x)$	Mean function of Gaussian Process
$C(x, x')$	Covariance function of Gaussian Process
$k(x, x)$	Kernel function of Gaussian Process Regression
$a(x)$	Acquisition function
I	Improvement function
EI	Expected improvement function
f^*	Best value know until current iteration in Bayesian Optimization
$\mu(x)$	Mean of normal distribution
$\sigma(x)$	Standard deviation of normal distribution
$\sigma^2(x)$	Variance of normal distribution
ϕ	Probability density function of normal distribution
Φ	Cumulative density function of normal distribution

n	Number of samples
N	Number of permutations in Permutation Feature Importance

1. INTRODUCTION

Throughout history, physicists, engineers and scientists have studied physical phenomena in pursuit of a better understanding of reality. As their comprehension of these phenomena grew, various models and analytical equations were developed. To this day, countless phenomena are still being studied and lack definitive mathematical formulations. In this context, experimental tests are employed as an approach to studying the physical properties of materials, and numerical models are developed to simulate these tests as realistically as possible. In another approach, data-driven models have increasingly gained prominence, due to the vast availability of data, the rise of low-cost computational power and advancements in Machine Learning (ML) techniques. Data-driven models can incorporate information from various tests, enabling predictions across a broader domain than equations derived from a single set of tests.

ML is a subfield of Artificial Intelligence (AI) that encompasses a wide range of algorithms and computational techniques where the computer is taught to perform tasks without being explicitly programmed to do so. Among these techniques are Artificial Neural Networks (ANNs), models composed of successive layers of interconnected neurons. The parameters of these models are weights and biases that apply linear transformations to the data, followed by non-linear activation functions, which grant these models the representational capacity needed to handle highly complex and non-linear problems. The traditional architecture of an ANN is called Multilayer Perceptron (MLP), which works well with multivariate numerical data that do not have time relation. In general, the aforementioned techniques can be applied to traditional regression and classification problems, but they can also perform tasks such as outlier detection, clustering and object detection, making them useful tools for a wide range of study areas, including Structural Engineering.

Obtaining better understanding of engineering problems and materials properties is fundamental to prevent accidents and ensure overall safety, as the design of structures rely on using equations and models for predicting relevant variables. A known accident caused by failure of metallic connections is the collapse of the Interstate 35W Highway Bridge in 2007 in Minneapolis, Minnesota, where 13 people died and 145 were injured (Figure 1.1). As reported by the National Transportation Safety Board (NTSB, 2008), its probable cause was the failure of the gusset plates (Figure 1.2) at the U10 nodes under a combination of significant increases

in the bridge's weight, caused by earlier structural alterations, along with the traffic and intense construction loads on the bridge on the day of the failure.

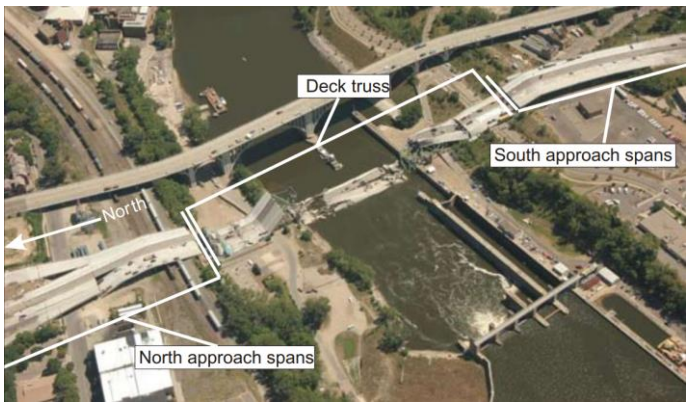


Figure 1.1. Interstate 35W Bridge after collapse (NTSB, 2008)

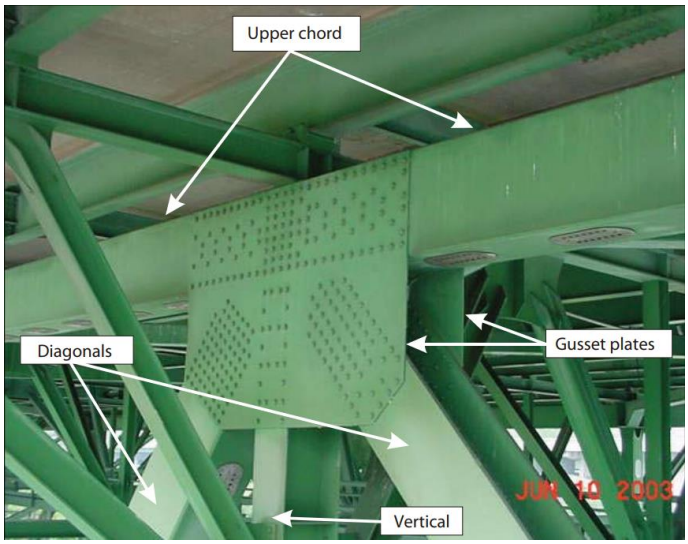


Figure 1.2. Gusset plates in five-member node on Interstate 35W Bridge (NTSB, 2008)

An example of complex physical phenomenon is shear lag, a nonuniform stress distribution across the cross-section of a structural member due to shear deformation. Generally, it happens when the element is unable to deform uniformly, for example when it is connected by only part of it, a common condition in bolted steel elements. Steel bolted angles under tension present a reduction in net section capacity due to shear lag and eccentricity of connection, because they are commonly connected by only one of its legs. This type of connection is very common in transmission towers, communication towers, and bridge diagonal members (QU; GUO; SUN, 2022; LI *et al.*, 2024), as shown in Figure 1.3. Many authors have studied the theme and

published their results, resulting in available data.

Figure 1.3. SBAs in transmission tower

Another complex phenomenon is the conduction of heat from solar radiation and environment to the bottom layers of concrete pavement. A combination of conduction, convection and radiation affect the temperature of all of its layers. The temperature of the concrete slab surface can be calculated with equations based on minimal inputs (SOLAIMANIAN; KENNEDY, 1993), being them hourly solar radiation and air temperature, as illustrated in Figure 1.4. With the assist of numerical simulations, it is possible to determine the temperature at the bottom of the concrete slab, comprising an interesting database for the development of surrogate data-driven models.

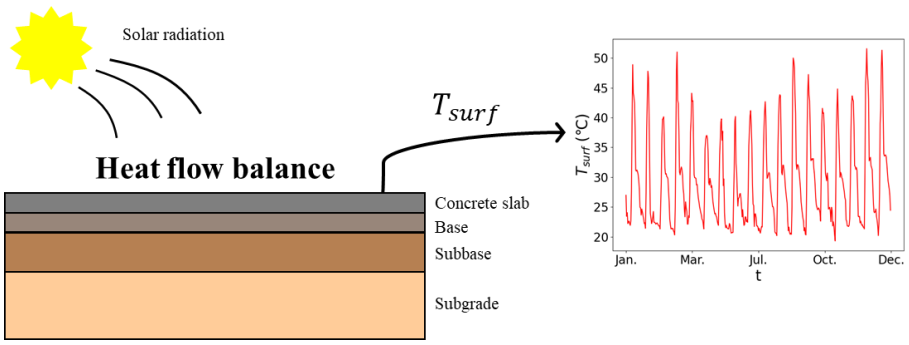


Figure 1.4. Illustration of surface temperature prediction in concrete pavements

In this context, ML techniques and Data-Driven Modeling (DDM) are tools capable of creating regression models that substitute experimental tests, numerical simulations, and standard equations. Such models can achieve high predictive accuracy, incorporate hybrid data, and significantly reduce financial and computational costs. Experimental tests require acquisition of material and availability of laboratory machines. Simulations based on Finite Element Method (FEM), for example, can take several days to run, while the training and optimization of MLPs take a few hours.

Aiming to demonstrate the effectiveness of ML-based DDM as a tool for the development of predictive models of reduced computational cost, two datasets were collected, comprising data from cold-formed steel bolted angles net section capacity and temperature variation on concrete pavement. ANN of MLP architecture were trained on these datasets. Bayesian Optimization (BO) and k-fold cross-validation were used to tune automatically hyperparameters of all the models. Regularization techniques era implemented to avoid overfitting, like early stopping and batch normalization. The accuracy of the models was evaluated with regression performance metrics on training and test sets, like Root Mean Squared Error and Mean Absolute Percentage Error. Predictive accuracy was visualized in scatter plots of predictions versus observed values, residuals and relative residuals. The regression models were interpreted applying feature importance techniques to extract the most important inputs for prediction of the output and also guide feature selection procedures, resulting in computationally cheaper and accurate enough models.

Building an accurate ANN model requires the adjustment of a lot of hyperparameters that define its architecture and learning process, such as learning rate, batch size, number of hidden layers and number of neurons in each hidden layer, thus requiring a hyperparameter tuning phase. This tuning process can be done with basic brute-force methods, like grid search and random search, metaheuristic algorithms, like Particle Swarm Optimization (PSO) and Genetic Algorithms (GA), or with sequential optimization strategies, such as Bayesian Optimization (BO). The choice of using BO for Hyperparameter Optimization (HPO) in the current work was due to its effectiveness in optimizing expensive black-box functions (SNOEK; LAROCHELLE; ADAMS, 2012; NGUYEN, 2019) and overall superiority to random search methods (TURNER *et al.*, 2021).

Another important step in ML modeling is Feature Selection (FS), the process of obtaining a relevant subset from an original feature set based on certain criterions (CAI *et al.*, 2018), as it improves predictive performance, provides faster and cost-effective predictors and facilitates data visualization and understanding (GUYON; ELISSEEFF, 2003). Very few studies that investigated BO directly applied to FS problems were encountered (GÖRMEZ, 2024; YANG, 2024). Based on these facts, the aforementioned methodology was used as baseline in comparisons with an algorithm that addresses FS and HPO on the same phase of ML model creation using BO. It was named SOFAH, Simultaneous Optimization of Feature Augmentation

and Hyperparameters, and it aims to improve models while adding information to the features set and reliably selecting a good performing subset.

Models based on baseline methodology and on SOFAH algorithm were developed for both mentioned collected datasets, comparing results. Also, public repository databases were used to compare results of baseline and SOFAH methodologies with the ones obtained by other authors.

1.1. MOTIVATION

Steel is one of the main structural materials in civil engineering, because of that it is important to understand its mechanical properties and failure mechanisms. Cold-formed steel is manufactured by bending, rolling or pressing steel, achieving light weight, high strength and ease of installation (ELLOBODY; YOUNG, 2005). In modern constructions, cold-formed steel is progressively being used as a primary structural component (XIAO *et al.*, 2022). Steel Bolted Angles (SBA) are L-shaped laminated steel elements that, when bolted and subjected to axial tension, may present net section failure, as one of some possible failure modes, caused by rupture of the section where the holes are located. These elements are widely used in various structures, like, power transmission towers, communication towers, and bridge diagonal members (QU; GUO; SUN, 2022).

SBA members with eccentric connections are common in civil engineering (BEHZADI-SOFIANI *et al.*, 2021) and are influenced by complex debilitating phenomena such as shear lag (KE *et al.*, 2018). Shear lag causes a nonuniform distribution of the tension stresses across the net section and, added to the effects of eccentricity, acts reducing angles tension capacity. In reason of that, the quantification of this reduction in efficiency of steel angles is a highly studied topic. Over the year, many researches were reported studying the theme and proposing new equations. Also, standards and codes around the world propose different equations for predicting the net section load capacity of SBAs.

Another material commonly used in civil engineering is concrete, that can compose rigid pavements, systems where the surface layers consists of a high stiffness concrete slab, followed by base, subbase and subgrade. Commonly, the slabs are directly exposed to solar radiation, thus conducting heat from the top to the bottom layers and presenting variation in volume and

length. Roadway pavements and the surrounding environment can be considered a microclimate system, since all layers are affected by their respective thermal properties and also climatic variables (CHEN; WANG; XIE, 2019). According to Mallick and El-Korchi (2023), the structural design of high-performance and long-lasting concrete pavements is based on the concept of limiting stresses and deformations to prevent excessive damage and deterioration of the system. One of the sources of stresses that need to be taken in consideration is stress due to temperature curling.

During the day, the surface temperature, at the top of the pavement increases due to solar radiation and high air temperatures. The bottom of the pavement, due to the relatively low thermal conductivity of concrete, takes longer to experience a temperature increase. As a result, expansion occurs more intensely at the top of the concrete than at its base, causing downward warping. At night, the situation reverses, with the surface of the slab cooling more rapidly than its base and causing contraction. The slab weight prevents curling, contraction or expansion, that is why tensile and compressive stresses are developed in the pavement (MALLICK; EL-KORCHI, 2023). The fact that temperature variates every single day puts these pavement systems under thousands of fatigue cycles over the decades. Combined with vehicular loading (CHEN; WANG; XIE, 2019), these stresses can damage the pavement and lead to cracking and deterioration, which reduces its performance over time. So, accurately evaluating variations in temperature in concrete pavements is fundamental to prevent fatigue failure and ensure the system does not reach the ultimate limit state.

The experimental analysis of temperature variations in rigid pavement, if done manually, can be extremely labor intensive, as it requires many temperature measurements approximately every hour. If done in automatic or semi-automatic ways, it becomes less labor intensive but more expensive. Experimental analysis of the net section capacity of SBAs depends on the occurrence of this specific mode of failure, while also requiring considerable infrastructure, like machines, steel angle samples, bolts, displacement measurement equipment, etc. Numerical modeling is a viable alternative that reduces many of the costs required for experimental tests, requiring only a computer of decent performance while also providing accurate results. Numerical techniques, such as FEM, demand good constitutive models that accurately describe the material responses to the loading conditions. So, considering that shear lag causes a complex stress state in SBAs under axial tension, numerical modeling of these elements may prove

challenging. High nonlinearity and complexity of the mentioned problems can be easily handled by powerful Machine Learning (ML) techniques, such as Artificial Neural Networks (ANN), being capable of provide cheaper and faster solutions, creating surrogate models that substitute equations, experiments and simulations. This is one of the main motivations of the present work and the reasoning behind the choice of using these tools.

Based on the considerable amount of experimental and numerical researches reported regarding both temperature in concrete pavement and SBAs under tensions, DDM is a great tool that makes good use of all the data already collected by other authors. This type of modeling, when based on ANN, has enough representation capacity to build excellent predictive models. Besides that, when aligned with statistical and sensitivity analyses, DDM can provide informative insights on the relations between inputs and outputs, indicating most relevant variables for example. Being one of the motivations of the current research the possibility of obtaining better understanding of the discussed physical phenomena.

The computational cost of ANN-based modeling is high during the hyperparameter optimization and training stages, but after one optimization cycle of hyperparameters and parameters, all their values are saved in files, resulting in cheap weight importation and very quick inference. Because of this, another one of the motivations of this research is to prove that ANN models can provide equal or greater accuracy in less time and computational effort than other numerical solutions.

When researching studies investigating the net section capacity of BSA connections, it was noted that there is a lack of studies addressing the topic using ANNs. As will be shown in chapter 0, there is only one single scientific paper (LI *et al*., 2024) that approaches the theme using ML techniques, but it studies specifically bearing capacity. This means that, so far, there is no published attempt of creating predictive models based on ML to study net section load capacity of these elements. As will also be shown in chapter 0, many studies already approached temperature prediction on pavements, but the vast majority were on asphalt pavement. Only a single study was found (HAN *et al.*, 2024) on the temperature prediction of concrete pavements using ML techniques, where the authors trained ML models to predict the total effective temperature difference based on FEM modeling results. Their aim was to evaluate the effects of curling on the concrete pavement. This lack of studies regarding the specific themes of the

current thesis was one very important motivation for it.

Building an accurate ANN model requires some steps that will be later discussed in the present work, but some of them are parameters optimization, commonly done with gradient-based optimizers, and hyperparameter tuning, commonly done with grid search or random search. Grid search is a brute-force-based search method that explores all configurations within a defined range. Random search is also a brute-force method, but instead of exploring all possibilities, it randomly chooses some of them from the total search space (RASCHKA; PATTERSON; NOLET, 2020). Although, these methods are exhaustive and inefficient, they are still widely used techniques, even in recent relevant scientific papers (BELETE; HUCHAIAH, 2022; SHAMS *et al.*, 2024). Another step comprehending ANN training is Feature Selection (FS), the process of selecting a subset of features reducing the original feature space, that is, performing dimensionality reduction. According to Cai *et al* (2018), FS consists of an optimization problem in which the perfect optimal solution can only be achieved by an exhaustive search and, because of that, researchers still use heuristic methods, like adding or removing features one at a time.

Based on this discussion, the final motivation of the present thesis is the demand for efficient algorithms capable of performing both hyperparameter optimization and Feature Selection simultaneously, specially using automatic searches, like the one performed in BO algorithm.

1.2. OBJECTIVES

The main goal of this work is to develop Machine Learning (ML) models with optimized hyperparameters and feature selection in regression models of selected and relevant engineering problems. It addresses the understanding of important physical properties using ANNs and Bayesian Optimization, creating accurate models that provide useful information about the analyzed variables. In this way, the research aims to demonstrate the effectiveness of this type of DDM in the study of physical phenomena.

The following specific objectives can be cited:

- Use of Bayesian optimization for hyperparameter optimization and feature selection;
- Quantify global uncertainties of the performance metrics for the developed models;

- Formulate and implement an algorithm, named SOFAH, that augments the feature space and simultaneously performs hyperparameters optimization and feature selection;

- Develop and evaluate ML-based data-driven models and strategies with different datasets of engineering problems from: public repository, experimental results, numerical simulations, and equations from standards;

- Compare the performance of baseline ML models with SOFAH-based models for different datasets;

- Verify if the present baseline methodology and SOFAH algorithm are competitive when compared to algorithms and models developed by other authors on the datasets from UCI ML repository;

- Collect and organize datasets from literature of cold-formed steel bolted angles, including various geometric properties of the angles, tensile strength of the steel material and ultimate capacity regarding net section failure;

- Create ANN models with data from numerical simulations and from standards' equations added to experimental data to obtain better predictive ML models of steel bolted angles net section strength;

- Collect and organize time series of climatological data to be used to predict surface temperature and differential temperature though the thickness of concrete pavements;

- Predict temperature in concrete pavements for posterior years and quantify the uncertainty present in these predictions.

1.3. STRUCTURE OF THE DISSERTATION

Chapter 0 addresses the foundations of the present work, encompassing literature review and theoretical and mathematical background of the main topics studied, detailing the concepts necessary to understand the present thesis. Chapter 3 discusses the methodology used in this work with respect to ML modeling, explaining baseline methodology and the implemented SOFAH algorithm. Chapter 4 approaches models created on UC Irvine ML Repository,

comparing baseline methodology, SOFAH-based models and other authors results. Chapter 0 presents the remaining methodology details of cold-formed steel bolted angles theme, specific to dataset and models developed, as well as results obtained, largely composed of tables and graphs. Chapter 6 shows the same, but for temperature variations of concrete pavement. Finally, chapter 7 presents the conclusions of the work and recommendations for future studies on the topic.

2. FOUNDATIONS

In this chapter, literature review is conducted by initially presenting previous studies and characterizing the state of the art on the topics addressed in this work. The main topics are: Bayesian Optimization (BO) applied to hyperparameters optimization and feature selection; steel angles connection tensile resistance; and temperature variations in concrete pavement. Then, theoretical and mathematical background is presented on generally the same topics, but now encompassing a lot more machine learning modeling concepts, not only BO, and approaching specifically SBA net section resistance. The background draws on various literature sources to explain in detail algorithms, techniques, methods, and mathematical formulations used in the research.

2.1. LITERATURE REVIEW

2.1.1. Bayesian optimization applied to feature selection and hyperparameter optimization

Bayesian Optimization (BO) originated with the work of Kushner (1964), Močkus (1975) and Zhilinskas (1975). Jones, Schonlau and Welch (1998) performed one of the main initial applications of BO, where they developed the efficient global optimization algorithm to solve engineering optimization problems, where the number of function evaluations is severely limited by time or cost. They used stochastic processes to create possible next search points that balance local and global search, that is, balancing exploration and exploitation. But it was only with the research of Snoek, Larochelle and Adams (2012) that BO received more attention within ML. The authors demonstrated that BO could be useful for deep neural networks, tuning nine hyperparameters of a three-layer convolutional network and achieving better accuracy on test set than when these hyperparameters were tuned by an expert.

In ML models, hyperparameters tuning can be formulated as an optimization problem where the objective function is an unknown derivative-free black-box function, so traditional optimization techniques like Newton method or gradient descent cannot be applied (WU *et al.*, 2019). As stated by Betrò (1991), Bayesian methods are very effective in solving this kind of problem, as they impose a probabilistic structure to information gained about the problem through function evaluations.

Since the mentioned studies, a lot more have been published indicating improvement in performance of ML models due to hyperparameter tuning with BO (VICTORIA; MARAGATHAM, 2021; GAO *et al*., 2021; LOEY; EL-SAPPAGH; MIRJALILI, 2022). But even with many studies regarding BO effectiveness, there is relatively little use of the technique in ML modeling. Bouthillier and Varoquaux (2020) showed that from all papers published at the Neural Information Processing Systems (NeurIPS) 2019 and the International Conference on Learning Representations (ICLR) 2020, although 80% of the NeurIPS papers and 88% of the ICLR papers tuned their hyperparameters, the large majority used manual tuning, random search, or grid search. Only 7% of the NeurIPS papers and 6% of the ICLR papers used a different method such as BO.

Motivated by the mentioned statistics, Turner *et al.* (2021) launched the black-box optimization challenge in 2020, a competition focused on tuning real-word ML tasks. The challenge showed that BO and similar methods are superior choices over random search and grid search for tuning hyperparameters of ML models. They found out that out of the 65 teams, 61 beat the baseline random search method and that all of the top-20 participants used some sort of surrogate-assisted optimization.

When performing the literature review on Bayesian Optimization (BO) applied to Feature Selection (FS) and Hyperparameters Optimization (HPO), many studies were found with BO and FS on their titles, but most of them did not apply BO to FS problems, the researches instead used FS methods and sequentially performed HPO with BO. Some of these works were Mate and Somai (2021) and Chaibi *et al.* (2022).

Mate and Somai (2021) created ML-based classification models for early-stage detection of breast cancer, aiming to increase survival rate and reduce treatment costs in patients and implementing FS and HPO algorithms. Initially, they use a voting system with six FS algorithms to determine optimal feature subset. Then, with this result, the models are trained and have their hyperparameters optimized with BO. Highlights of the work primarily are the high accuracy obtained by their best model, of 97%, the voting system for FS and that a high quantity of ML techniques were implemented, 15 in total, all with hyperparameters optimized. The methodology implemented by the authors consist of a quasi-standard methodology for ML modeling, since it follows data preprocessing, feature selection, model building and model

evaluation, with FS being performed in a completely separated step from HPO. Chaibi *et al.* (2022) used five ML techniques to develop regression models that predict daily global solar radiation based on FS and HPO algorithms. They first evaluate feature importance based on random forests, then, with this result, perform FS. With optimal inputs defined, the models are trained and have their hyperparameters optimized with BO. In practice, this method implemented by the authors consist of a standard two-step FS and HPO, performed separated from each other.

Just a few studies were found applying BO directly as a tool for FS problems. Görmez (2024) directly used BO for FS in classification models, using 9 distinct ML techniques in two datasets available in sklearn library. He states that his methodology is new and based on the optimization of a vector of size M x 1, where M is the number of features, consisting of 0 and 1 values, 0 indicating that the corresponding feature is not included and 1 that it is included. Feature selection, model training, performance score calculation, and expected improvement calculation steps are repeated a predetermined number of times, for the features that yield highest performance score to be chosen. Despite the novel methodology, the author used all hyperparameters values as their default in the sklearn library. Yang, Liu and Wen (2024) investigated embedded-based FS algorithms that have hyperparameters, extreme gradient boosting, Lasso and elastic net, optimizing their values with BO. They analyzed high-dimensional molecular data, established 100 iterations for BO and pre-specified the number of selected features to 100, 200, 500 and 1000. The authors concluded that BO generally improved recall rate but, depending on the complexity of hyperparameters space, results may vary.

2.1.2. Steel angles connection tensile resistance

The net section reduction coefficient U, has been studied for over 60 years (CHESSON JR., 1959), with one of its first formulations being the equation $U = 1 - \bar{x}/L$, proposed by Munse and Chesson Jr. (1963). Many years later, LaBoube and Yu (1996) proposed a new equation based on the results of Holcomb, LaBoube and Yu (1995) and the original equation by Chesson (1959). Yip and Cheng (2000) conducted experiments on 12 angles, and Paula *et al.* (2008) performed 66 experimental tests, both proposing new equations for the reduction factor due to shear lag. Teh and Gilbert (2013) analyzed the accuracy of the equations presented by the American and Australasian standards, highlighting the high accuracy of the equation proposed

in their work, an equation with a format very similar to the one currently presented in the American Iron and Steel Institute (2016) standards.

Ke *et al.* (2018) investigated the shear lag effect on ultimate tensile capacity of high strength steel angles with bolted and welded connections, evaluating the effects of steel grade, connection length and out-of-plane eccentricity. Jiang and Zhao (2022) presents experimental and numerical studies of stainless angles connected by one leg and failing by net section fracture, comparing results with European and American codes. Paula *et al.* (2008) performed an experimental study on cold-formed steel angles fastened with bolts and under tension, focusing on the effects of the shear lag phenomenon and its consequent reduction in net section capacity, proposing a new equation for this reduction factor. Fleitas *et al.* (2020) conducted a parametric numerical study on the same theme and proposed two similar equations with different coefficients, for angles with one and two bolt lines.

Machine Learning (ML) has been widely applied in civil and structural engineering, many important mechanical properties have proved to be predictable with DDM, like concrete compressive strength (SILVA *et al.*, 2023; CHOU *et al.*, 2014) and compressive and flexural strengths of steel fiber-reinforced concrete (KANG; YOO; GUPTA, 2021). Moreover, ML have been used predict other variables that are not mechanical properties of structural materials, predicting failure probabilities (LIMA; EVANGELISTA JR.; SOARES, 2023), quantifying uncertainties (EVANGELISTA JR.; ALMEIDA, 2021), and consequently assisting structural reliability analyses.

Sarothi *et al.* (2022a) used eleven ML techniques as an approach to studying the bearing capacity of double shear bolted connections and performed grid search tuning, an exhaustive and expensive technique, with 10-fold cross-validation to select optimal hyperparameters values. They developed a user interface for generalized data-driven design of bolted connections, irrespective to the region, steel grade, bolt arrangements and type of failure. The same authors (2022b) implemented ten ML approaches to classify the failure mode of double shear bolted steel connections, again using grid search with 10-fold cross-validation. They developed a new user interface that identifies failure type of double shear bolted connections, independent of specific connection geometric and material configuration. In both researches, only flat sheet connections were studied, in opposition to the present work, that addresses SBA,

elements which load capacity is highly affected by shear lag and eccentricity of connection.

Xiao *et al.* (2022) also used ML methods applied to cold-formed steel members, but to study the loading capacity prediction and optimization of built-up section columns, using Shapley Additive Explanations to investigate feature importance and dependency. Li *et al.* (2024) conducted the only published study that approaches SBA resistance with ML techniques, however they analyzed bearing capacity, as the present work analyzes net section capacity. The authors created a new database of SBA with one and two bolts disposed in a single line, comprising a total of 175 instances, 51 experimental and 124 numerical. The authors established a convenient Graphic User Interface based on their two best performing models. So, no work from SBA's literature addresses net section load capacity with ML techniques.

2.1.3. Temperature variation in concrete pavements

The structural design of concrete pavements has been studied for many years, with researches published since the early 20th century. Westergaard (1925) was the first one to develop a rational theory capable of calculating stresses in concrete pavement slabs. He assumed the slab to act as a homogeneous isotropic elastic solid in equilibrium and the reactions of the subgrade to be vertical only and proportional to the deflections of the slab. With these assumptions, the problem was reduced to a problem of mathematical theory of elasticity. Westergaard (1926) developed further his theoretical analysis, providing formulas, charts and tables which comprise a convenient method that addressed critical stresses in existing pavements and allowed for reduction in thickness of pavements, thus resulting in more efficient highway engineering. Westergaard (1927) observed that cracks developed in new pavement before any load was put on it, showing how important were the stresses due to temperature variations. The author investigated a specific case of quick changes of temperature, like the change from a cold night to a hot day, analyzing concrete slab curling and solving equations to obtain stress values.

Teller and Sutherland (1935) published a long and detailed report on extensive investigation on the structural action of concrete pavement slabs. In part 2 out of 5 of their report, the authors described their results regarding observed effects of variations in temperature and moisture on the size, shape and stress resistance of concrete pavement slabs. They reached many important conclusions for the time, such as: average pavement temperature annual change of about 26.67 ^{o}C, maximum temperature differentials in hot afternoons of early summer, cyclic variation in

slab length from temperature changes, experimental verification of the thermal coefficient of expansion of the concrete, etc.

Bradbury (1938) researched the structural design of reinforced concrete pavements, developing the theoretical background of the subject. He emphasized the application of rational stress determinations to the structural analysis of concrete pavements as a useful analytical tool for evaluating the merits of slab thickness, joint spacing, reinforcement and other features affecting structural strength.

Regarding analytical models, earliest studies on analytical solutions of pavement temperature were developed by Barber (1957), deriving a solution based on weather data. He applied a thermal diffusion theory to a semi-infinite mass, the pavement, in contact with air. Some of the limitations of his method are that it uses total daily radiation and it does not consider latitude effects. Based on this, Solaimanian and Kennedy (1993) proposed a simple analytical method to predict maximum temperature on pavement surface based on an equilibrium state and using minimal inputs, only maximum air temperature and maximum hourly direct solar radiation, as that translates to knowing latitude and air temperature. Their results were compared to measured pavement surface temperatures, achieving a difference of 4ºC in 96% of the measurements and a difference of 3ºC in 83%.

After that, a series of relevant studies were published by mostly the same authors, all solving the partial differential equation for heat conduction through analytical methods (CHEN; WANG; XIE, 2019) as some of the methods used were Hankel integral transform (WANG; ROESLER; GUO, 2009), separation of variables (WANG; ROESLER, 2014; WANG D., 2015), Duhamel's principle (WANG D., 2015), Laplace transform and inverse Laplace transform (WANG D., 2012; WANG; ROESLER, 2012), and eigenfunction expansion technique (WANG D., 2016). All these analytical solutions approached the prediction of temperature profiles in multilayered pavement systems or of just a single layer pavement.

Machine Learning has been in used in a variety of studies to predict performance-related properties of pavement, as Piryonesi and El-Diraby (2021) implemented several algorithms to predict a single index, pavement condition index, based on climate data, considering both a current climate scenario and a climate change scenario in 2098. Zeiada *et al.* (2020) investigated pavement design factors that most influence its performance in warm regions using five ML

algorithms and analyzing many structure, climate and traffic related inputs to predict the international roughness index. Both mentioned researches studied the performance of asphalt pavement.

A lot of recent studies were found regarding ML applications on the prediction of temperature: in pavement surface (TABRIZI *et al*., 2021), integrating convolutional neural networks with long short-term memory networks; at different depths (NOJUMI *et al.*, 2022), implementing five usual ML algorithms; both in pavement surface and varying depths (MILAD *et al.*, 2021), with a hybrid ML model that used Markov chain Monte Carlo and Random Forest; in base and subgrade layers (HUANG *et al*., 2023), using few inputs and five usual ML algorithms. Even though all these researches were found, all of them are specific for asphalt pavement.

From the performed literature review, it can be concluded that there is significant lack of studies regarding temperature prediction in concrete pavements using ML techniques, as compared to flexible pavement, almost no studies were found on the theme. Han *et al*. (2024) very recently used ML in a study that evaluated the degree of total curling specifically in concrete pavement. The author simulated falling weight deflectometer testing in 3D FEM models, thus obtaining a large dataset to train models with ML techniques that surrogate the computationally demanding FEM models. He achieved accurate models capable of predicting the total effective temperature difference to quantify the effects of curling on the concrete pavement.

2.2. THEORETICAL AND MATHEMATICAL BACKGROUND

2.2.1. Machine learning modeling

This section focuses exclusively on aspects related to ML modeling, explaining most of the background of the techniques used in the research, with various references of the literature of Bishop (2006) and Goodfellow, Bengio and Courville (2016). Also, the methodology related to such modeling is fully addressed, detailing the algorithm developed by the present author, which was applied to all datasets studied in the research.

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that encompasses a wide range of algorithms and computational techniques where the computer is taught to perform tasks without being explicitly programmed to do so. These algorithms learn through

“experience”, that is, they always learn through data, thus always involving DDM. In this context, the first step in ML modeling is the acquisition and preparation of a high-quality dataset.

Data gathering for ML modeling is most of the times an iterative process that follows the general steps of Figure 2.1. Starting with the acquisition of a considerable number of samples that need to go through statistical analysis, using descriptive statistics to measure distribution, variability and central tendency of each variable. The statistical analysis enables a more efficient process of feature engineering, easing the detection of outliers and near-zero or zero variance features. Some data handling algorithms and libraries requires non-missing values, thus requiring solutions to this type of values, either by removing samples with missing values or by imputing them on the training set, usually based on mean or median for quantitative variables and most frequent value for categorical features. Feature transforming methods can be used to add useful information to the dataset and consequently to the model, by performing, for example, square, cube, reciprocal and logarithm transforms or by performing operations such as multiplication or division between two variables. The dataset is split into training, validation, and test sets and undergoes normalization before finally being used in model development. If the model's accuracy is not satisfactory, there is a wide range of solutions to be employed, one of which is acquiring more data, starting the steps all over again.

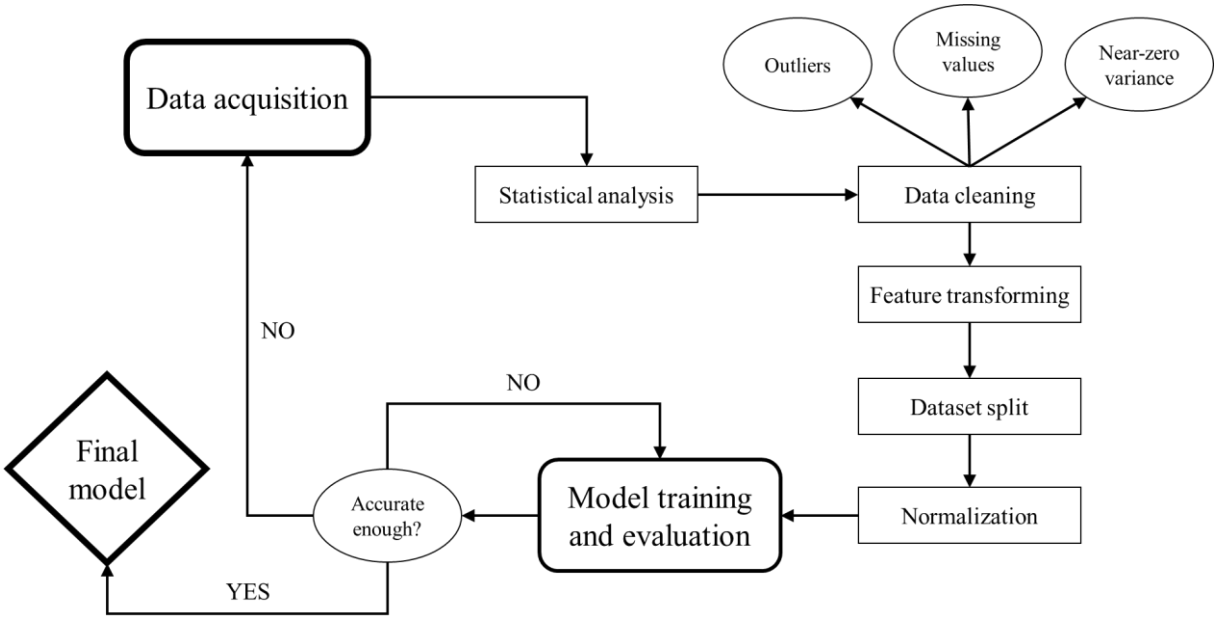


Figure 2.1. High-quality dataset construction steps

The number of samples obtained directly affects the quality of the final model, as a higher number allows training models with a greater number of parameters without overfitting, an undesirable behavior that will be explained in detail in following paragraphs. Additionally, the broader the range of the variables, the larger the domain in which the regression model performs well, without the need for extrapolation.

Train-validation-test split does not have stablished best proportions to be performed, varying a lot in the literature. However, it is an important step that defines different sets of samples to be used in each of these phases, as normalization and imputation of missing values need to be done based on training set in order to not leak information from test or validation. The training phase is when model's parameters $\boldsymbol{\theta}$ are optimized minimizing a cost function, usually Mean Squared Error (MSE) in regression problems, learning iteratively how to predict the target variable y, where model's prediction $\hat{y}$ are given by the set of parameters $\boldsymbol{\theta}$. The objective of this phase is to create a model that captured the patterns present in input variables $\boldsymbol{x}$, resulting in the capability of predicting accurately y given an unseen sample of inputs $\boldsymbol{x}$.

Validation and test are very similar phases that aim to evaluate model's performance performing predictions on unseen data. The only thing that differentiate them is that validation results are used for model calibration, as in test model's final performance is measured. Validation phase ensures that the model is not only learning training set patterns but also maintaining adequate generalization capability, while also being useful to adjust model's hyperparameters.

When data is limited, cross-validation techniques present much better use of data, as they allow the usage of subsets of the training set as validation set, instead of requiring a separated set exclusive to the validation task and another for test. In this context, k-fold cross-validation divides the data in a number k of folds, with one of them being selected as the validation set and the rest as training set, this process is repeated k times always selecting a different fold for validation. The algorithm is illustrated in Figure 2.2.

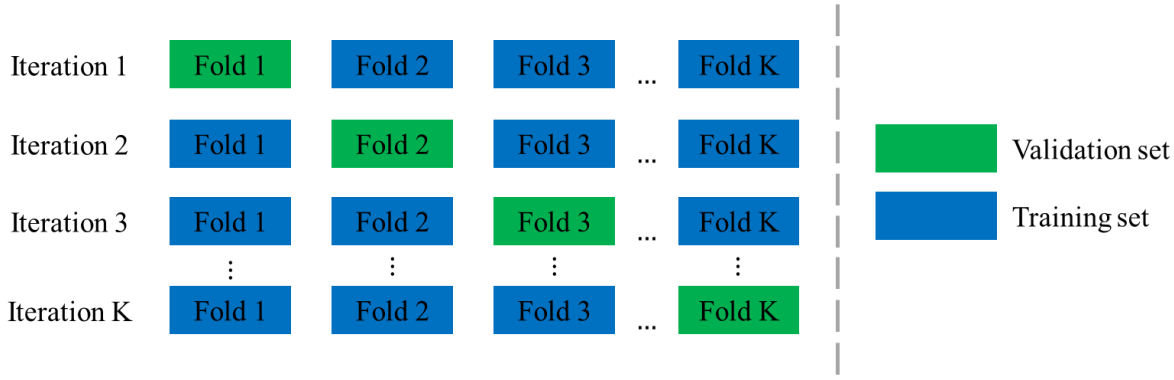


Figure 2.2. K-fold cross-validation algorithm

The ideal model achieves a good balance between two sources of error in predictive models, bias and variance, that when one is high, the other is low, and vice versa. Bias refers to errors caused by simplistic assumptions in the learning algorithm, so high bias models fail to capture underlying patterns in the data, resulting in underfitting. Variance refers to errors caused by model's sensitivity to small fluctuations in the training set, so high variance models adjust excessively to the training data, learning even its noise and leading to overfitting. In mathematical terms, considering that $\boldsymbol{\theta}$ represents the true values of model's parameters and that $\widehat{\boldsymbol{\theta}}$ represents a point estimate of parameters $\boldsymbol{\theta}$, let $\{x^{(1)}, \dots, x^{(m)}\}$ be a set of m independent and identically distributed (i.i.d.) data points, a point estimator is any function of the data:

$$\widehat{\boldsymbol{\theta}}_m = g\left(x^{(1)}, \dots, x^{(m)}\right) \tag{2.1}$$

The bias of an estimator is defined as:

$$\text{Bias}\left(\widehat{\boldsymbol{\theta}}_m\right) = E\left[\widehat{\boldsymbol{\theta}}_m\right] - \boldsymbol{\theta} \tag{2.2}$$

$$\text{Bias}\left(\widehat{\boldsymbol{\theta}}_m\right)^2 = E\left[\widehat{\boldsymbol{\theta}}_m\right]^2 - 2\boldsymbol{\theta}E\left[\widehat{\boldsymbol{\theta}}_m\right] + \boldsymbol{\theta}^2 \tag{2.3}$$

The variance of an estimator is defined as simply

$$\text{Var}\left(\widehat{\boldsymbol{\theta}}_m\right) = E\left[\left(\widehat{\boldsymbol{\theta}}_m\right)^2\right] - E\left[\widehat{\boldsymbol{\theta}}_m\right]^2 \tag{2.4}$$

The MSE of the estimates:

$$\text{MSE} = E\left[\left(\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\right)^2\right] \tag{2.5}$$

$$\text{MSE} = E\left[\left(\widehat{\boldsymbol{\theta}}_m\right)^2\right] - 2\boldsymbol{\theta}E\left[\widehat{\boldsymbol{\theta}}_m\right] + \boldsymbol{\theta^2} \tag{2.6}$$

$$\text{MSE} = E\left[\left(\widehat{\boldsymbol{\theta}}_m\right)^2\right] - 2\boldsymbol{\theta}E\left[\widehat{\boldsymbol{\theta}}_m\right] + \boldsymbol{\theta^2} + E\left[\widehat{\boldsymbol{\theta}}_m\right]^2 - E\left[\widehat{\boldsymbol{\theta}}_m\right]^2 \tag{2.7}$$

$$\text{MSE} = E\left[\widehat{\boldsymbol{\theta}}_m\right]^2 - 2\boldsymbol{\theta}E\left[\widehat{\boldsymbol{\theta}}_m\right] + \boldsymbol{\theta^2} + E\left[\left(\widehat{\boldsymbol{\theta}}_m\right)^2\right] - E\left[\widehat{\boldsymbol{\theta}}_m\right]^2 \tag{2.8}$$

$$\text{MSE} = \text{Bias}\left(\widehat{\boldsymbol{\theta}}_m\right)^2 + \text{Var}\left(\widehat{\boldsymbol{\theta}}_m\right) \tag{2.9}$$

Based on Equation (2.9), it is clear that the cost function MSE incorporates both bias and variance, thus reducing MSE tends to also reduce bias and variance. Figure 2.3 illustrates the bias-variance tradeoff in function of model's capacity, since an increase in capacity is equivalent to more parameters and consequently more variance, that is, ease of overfitting. This concept is fundamental to ML modeling, because it ensures that, when using optimal model capacity, an optimized MSE value translates to low bias and low variance.

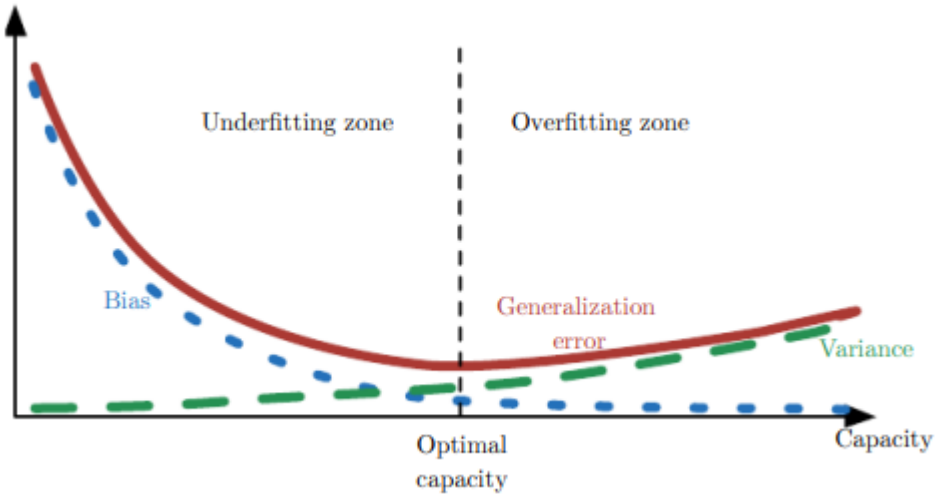


Figure 2.3. Bias-variance tradeoff and model capacity (GOODFELLOW; BENGIO AND COURVILLE, 2016)

2.2.1.1.Artificial Neural Networks

Artificial Neural Networks (ANN) are widely one of the main computational models with structure inspired by human brains, presenting neurons disposed in layers connected to each other. The Multilayer Perceptron (MLP) is a fully connected feedforward ANN, where there are at least three layers (Figure 2.4), one input layer, one or more hidden layers and one output layer. These layers are densely connected, with all its neurons connected with all neurons of the next layer, opposing sparsely connected layers.

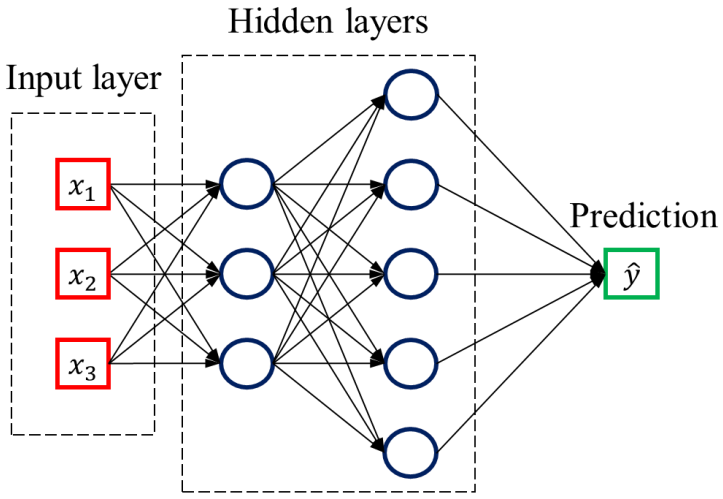


Figure 2.4. Multilayer Perceptron architecture

Every connection in the network has associated weight (w) and bias (b), which form the parameters of the model, calculated usually with gradient-based optimizers in the training phase. In feedforward networks, the information goes through the model in one direction only, from input to output, with no cycle among the neurons. The output of the input layer, i.e., the input of the first hidden layer, is given by

$$\hat{\boldsymbol{y}} = f_a(\boldsymbol{XW} + \boldsymbol{b}) \tag{2.10}$$

where $\boldsymbol{X}$ is a matrix with each row representing an instance and each column a feature, f_a is the nonlinear activation function, $\boldsymbol{W}$ is the weight matrix and $\boldsymbol{b}$ the bias vector.

Gradient-based optimization algorithms refer to minimizing the loss function, usually MSE when approaching regression problems, by changing parameters' values in small steps in the opposite direction of gradient. The size of these steps is defined by a hyperparameter α called learning rate, that can be fixed or, in modern algorithms, adaptive. This process is only viable because of the backpropagation algorithm (RUMELHART; HINTON; WILLIAMS, 1986), that presents an efficient technique of passing information forward through the network to produce a scalar loss and then backward to compute inexpensively the gradients, applying recursively the chain rule of calculus. Also, because of fundamental contributions of Bishop (1995, 2006).

Activation functions, f_a, are functions that provide nonlinearity between layers, enabling ANNs to approach complex nonlinear problems. These functions are commonly chosen as logistic, hyperbolic tangent or Rectified Linear Unit (ReLU). ReLU is the actual default

recommendation for most feedforward ANNs (BISHOP; BISHOP, 2023), because of its nearly linear format, preserving properties that make linear models generalize well and optimize easily with gradient-based techniques. Extremely relevant papers and authors have showed that ReLU makes training faster and more stable (NAIR; HINTON, 2012; KRIZHEVSKY; SUTSKEVER; HINTON, 2012), while also being cheap to evaluate (BISHOP; BISHOP, 2023).

An important element of ANNs training is the initialization of its parameters, determining whether learning does converge or not, quickly or slowly and to a point with high or low cost. Common practice is to initialize the weights of a fully connected layer sampling each weight from a uniform or normal distribution of parameters defined by the number of inputs and outputs of the given layer. He *et al*. (2015) proposed that, for the variance of the outputs of each layer to be equal to the variance of its inputs, weights should be initialized sampling them from a zero-mean Gaussian distribution of standard deviation $\sqrt{2/n_l}$, being n_l the number of inputs of the layer l if preserving the magnitude of the weights in the forward pass.

Regarding optimization algorithms with adaptive learning rates, Adam (KINGMA; BA, 2014), derived from Adaptive moment estimation, is a method for efficient stochastic optimization. It computes individual adaptive learning rates for different parameters from estimates of first and second moments of the gradients. The algorithm keeps track of an Exponentially Weighted Moving Averages (EWMA) of the gradient and an EWMA of the squared gradients. The steps of the algorithm are:

$$\boldsymbol{m} \leftarrow \boldsymbol{\gamma_1 m} + (\mathbf{1} - \boldsymbol{\gamma_1})\boldsymbol{\nabla_\theta J(\theta)}$$

$$\boldsymbol{v} \leftarrow \gamma_2 \boldsymbol{v} + (1 - \gamma_2)\left(\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})\right)^2$$

$$\widehat{\boldsymbol{m}} \leftarrow \frac{\boldsymbol{m}}{1 - {\gamma_1}^t} \qquad \text{2.11(a-e)}$$

$$\widehat{\boldsymbol{v}} \leftarrow \frac{\boldsymbol{v}}{1 - {\gamma_2}^t}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \alpha \frac{\widehat{\boldsymbol{m}}}{\sqrt{\widehat{\boldsymbol{v}} + \delta}}$$

where γ_1 and γ_2 are the exponential decay rates for the moment estimates, m and v are the biased first and second moment estimates, $\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta})$ is the gradient of the loss function $J(\boldsymbol{\theta})$ with

respect to parameters $\boldsymbol{\theta}$, $\hat{\boldsymbol{m}}$ and $\hat{\boldsymbol{v}}$ are the bias-corrected first and second moment estimates, δ is a small constant for numerical stability and α is the learning rate. Reddi, Kale and Kumar (2019) have shown that Adam does not always converge to the optimal solution and proposed AMSGrad, a variant of Adam with guaranteed convergence. The variant provides a new exponential moving average, taking the maximum of all second moment estimates until present time step and normalizing the running average of the gradient.

2.2.1.2.Regularization

Regularization techniques are methods that assist the development of models of great generalization capability, preventing overfitting, situation where there is excessive learning of the patterns in training set, capturing also noise and outliers, thus performing poorly on unseen data. Some of these techniques are:

- Early Stopping, consists in stopping model training when validation error is not improving after a given number of consecutive iterations, also called patience, thus preventing the model to perform excessive training;

- Weight Decay, modifying the loss function in training by adding a criterion that penalizes large weight values based on squared L2 norm and on a hyperparameter λ that controls the strength of this preference for smaller weights;

- Batch Normalization (IOFFE; SZEGEDY, 2015), is a method of adaptive reparametrization that speeds up training and improves model performance. It normalizes the inputs of each layer to have mean 0 and variance 1, then, in order to maintain representation capacity of the model, it learns parameters γ and β in training to respectively scale and shift the result.

2.2.1.3.Bayesian Optimization

Bayesian Optimization (BO) is a black-box optimization algorithm that relies on a probabilistic surrogate model for the objective function, providing a measure of uncertainty. Based on this surrogate model, usually Gaussian Processes (GP), an acquisition function determines the most promising point to be evaluated next (TURNER *et al*., 2021). In general, the technique works well in optimization problems where function evaluations are expensive, in a way that it is

desirable to spend computational time making good choices on where to perform new evaluations in order to find the best parameters (SNOEK; LAROCHELLE; ADAMS, 2012).

The technique requires user-defined ranges for the search space for a given number of iterations or until convergence. These facts represent disadvantages of the algorithm, as they require prior knowledge or multiple configuration tests, resulting in large spaces of inefficient optimization or small spaces that do not contain the optimum point (NGUYEN, 2019). On the other hand, once the ranges are defined reasonably and either a sufficient number of iterations or a satisfactory convergence criterion are established, the algorithm automatically finds optimal solutions.

In Figure 2.5, it is shown generally how the BO algorithm works, where the bottom of each plot shows the acquisition function and the top shows a GP approximation of the objective function. The figure starts at iteration 2, where 2 observations from the function were evaluated, represented by the black points. Based on these observations, mean and standard deviation of the GP are evaluated, approximating the objective function and providing the blue area of posterior distribution uncertainty. Then, with assist of GP regression, the acquisition function is evaluated and its maximum is selected as the next point to be evaluated, providing a new observation that will be included in a new GP approximation that start the iterative process again. So, even if the new observation is not a point of minimum, it still added information to the GP, aiding the search for the minimum on next iterations.

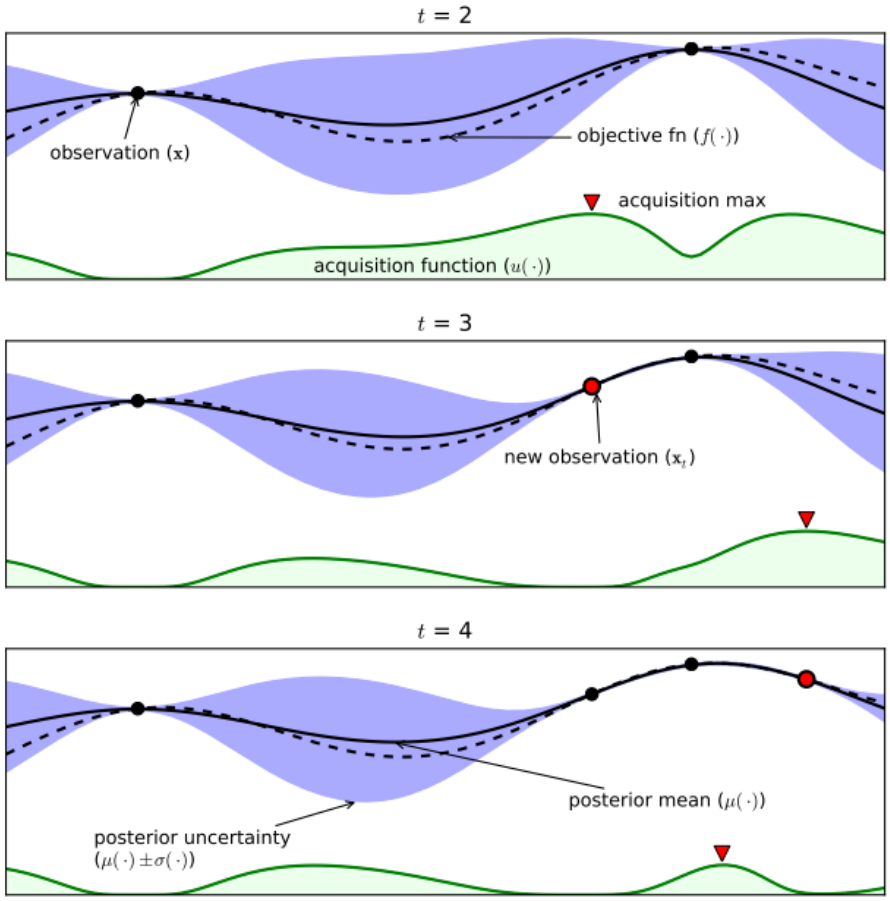


Figure 2.5. Illustration of 3 iterations of Bayesian Optimization algorithm (BROCHU; CORA; FREITAS, 2010)

In Figure 2.6, a pseudocode of the BO algorithm can be seen, where the inputs are: $\boldsymbol{S}$, the search space; x_0, the arbitrarily chosen initial point from the search space $\boldsymbol{S}$; N_ITERS, the first criterion of convergence, a maximum number of iterations for the algorithm; and PATIENCE, a second criterion of convergence, a limit of consecutive iterations of no improvement in the point of minimum. The functions are: $objective(x_i)$, an evaluation of the objective function at point x_i; $random(\boldsymbol{S})$, the sampling of a random point of search space S; $acq\_func(\boldsymbol{x})$, the acquisition function over all observed points $\boldsymbol{x}$; and $GP(\boldsymbol{x}, \boldsymbol{y})$, a new fit of Gaussian Process (GP) with all observed points $(\boldsymbol{x}, \boldsymbol{y})$. In each iteration, at x_i, the objective function is evaluated, y_i, and both coordinates are appended to lists that have all observed values. In the first 10 iterations, random points x_i are used, then, in each subsequent iteration, a new GP is fit on all observed points and new points x_i are selected based on the maximum of the acquisition function. The value returned by the algorithm is the last point of minimum min_x, selected from search space $\boldsymbol{S}$.

inputs: S, x_0, N_ITERS , PATIENCE

functions: $objective(x_i)$, $random(S)$, $acq\_func(x)$, $GP(x, y)$

$\text{LOSS\_MIN} = \infty$

$x, y = list(), list()$

for $i = 0: \text{N\_ITERS} - 1$:

    $y_i = objective(x_i)$

    $x, y = x.append(x_i), y.append(y_i)$

    if $y_i < \text{LOSS\_MIN}$:

        $x\_min = x_i$

        $\text{LOSS\_MIN} = y_i$

        $\text{TOL} = \text{PATIENCE}$

    else:

        $\text{TOL} = \text{TOL} - 1$

        if $\text{TOL} = 0$:

            break

    if $i < 10$:

        $x_{i+1} = random(S)$

    else:

        $GP(x, y)$

        $x_{i+1} = \text{argmax}\, acq\_func(x)$

return x_min

Figure 2.6. Pseudocode of Bayesian optimization algorithm

According to Snoek, Larochelle and Adams (2012), as in other optimization problems, the desired solution is the minimum of a function $f(\boldsymbol{x})$ on some bounded set χ, taken to be a subset of $\mathbb{R}^D$. BO, in particular, constructs a probabilistic model for $f(\boldsymbol{x})$ to use all of the information available from previous evaluations of the function. The computational cost of determining the next point to evaluate is easy to justify due to the cost of performing each evaluation. Selecting a GP prior distribution, taken to be of the form $f: \chi \to \mathbb{R}$, it is defined by the property that any finite set of N points $\{x_n \in \chi\}_{n=1}^{N}$ induces a multivariate Gaussian distribution on $\mathbb{R}^N$.

Based on the paper of William and Rassmussen (1996), on their book (RASSMUSSEN;

WILLIAMS, 2006) of Gaussian Processes for Machine Learning, and the literature of Bishop (2006), GPs are a collection of random variables, any finite number of which have a joint Gaussian distribution, and they are completely specified by mean and covariance functions. Mean function $m(\boldsymbol{x})$ and covariance function $C(\boldsymbol{x}, \boldsymbol{x}')$ of a real process $f(\boldsymbol{x})$ is defined as

$$m(\boldsymbol{x}) = \mathrm{E}[\mathrm{f}(\mathbf{x})] \tag{2.12}$$

$$C(\boldsymbol{x}, \boldsymbol{x}') = \mathrm{E}[\big(\mathrm{f}(\mathbf{x}) - \mathrm{m}(\mathbf{x})\big)\big(\mathrm{f}(\mathbf{x}') - \mathrm{m}(\mathbf{x}')\big)] \tag{2.13}$$

and the GP can be written as

$$f(\boldsymbol{x}) \sim GP\big(m(\boldsymbol{x}), C(\boldsymbol{x}, \boldsymbol{x}')\big) \tag{2.14}$$

For simplicity, mean function, $m(\boldsymbol{x})$ is taken to be zero. In order to apply GP models to the problem of regression, the random noise ϵ_n on the observed target values needs to be considered. The noise process is considered to have a Gaussian distribution and to be based on a hyperparameter β_G that represents the precision of the noise.

The objective is to predict target variables for new inputs given a set of training data, So, supposing that the training set is composed of targets $\boldsymbol{y}_N = (y_1, \dots, y_N)^T$ that corresponds to input values $x_1, \dots, x_N$, the objective is to predict y_{N+1} given x_{N+1}. This requires evaluating the predictive distribution $p(y_{N+1}|\boldsymbol{y}_N)$, given by a Gaussian with mean and covariance

$$m(x_{N+1}) = \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{y} \tag{2.15}$$

$$\sigma^2(x_{N+1}) = c - \boldsymbol{k}^T \boldsymbol{C}_N^{-1} \boldsymbol{k} \tag{2.16}$$

where vector $\boldsymbol{k}$ has elements $k(x_n, x_{N+1})$ for $n = 1, \dots, N$, scalar $c = k(x_{N+1}, x_{N+1}) + \beta_G^{-1}$ and C_N is the $N \times N$ covariance matrix $C(x_n, x_m) = k(x_n, x_m) + \beta_G^{-1}\delta_{nm}$ for $n, m = 1, \dots, N$. Note that δ_{nm} is a Kronecker delta which is 1 if and only if $n = m$ and 0 otherwise. The term $k(x_n, x_m)$ is a kernel function for GP regression, that can assume many functions, for example the Matérn class of covariance functions and the specific Matern 5/2, respectively

$$k_{Matern}(r) = \frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}r}{l}\right)^{v} K_v\left(\frac{\sqrt{2v}r}{l}\right) \tag{2.17}$$

$$k_{v=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2}\right)\exp\left(-\frac{\sqrt{5}r}{l}\right) \tag{2.18}$$

With Gaussian Processes and Gaussian Process Regression defined, it is clearer how the techniques quantify the uncertainty of unobserved areas based on previous observations and provide a Bayesian posterior probability distribution. They combine prior information about the unknown function with sample information, to acquire posterior information of the function distribution with Bayesian formula. Optimal value location is deduced based on this posterior information (BETRÒ, 1991). Although, this optimal value is based on an acquisition function, not on the original function, because the acquisition function $a(x)$ is inexpensive and of known structure, thus it provides the next query point, i.e., next point to be evaluated in the original function.

A usual and popular acquisition function is Expected Improvement (*EI*). As explained by Brochu, Cora and Freitas (2010), the improvement function $I(x)$ is given by

$$I(\boldsymbol{x}) = \max(0, f_{N+1}(\boldsymbol{x}) - f^*) \tag{2.19}$$

where $f_{N+1}(\boldsymbol{x})$ is the evaluation of a given point $\boldsymbol{x}$ and f^* is the best value known thus far. So, choosing a next point $\boldsymbol{x}$ to be evaluate, it returns the difference between its evaluation $f(x)$ and f^* if $f(x) > f^*$, otherwise, returns zero. The new query point $\boldsymbol{x}$ is found by maximizing the expected improvement:

$$\boldsymbol{x} = \operatorname{argmax} \mathrm{E}[\max(0, f_{N+1}(\boldsymbol{x}) - f^*)|\mathcal{D}_N] \tag{2.20}$$

where $\mathcal{D}_N$ is the prior distribution. The likelihood of improvement I on a normal posterior distribution characterized by $\mu(x)$ and $\sigma^2(x)$ can be computed from the normal probability density function ϕ

$$\phi = \frac{1}{\sqrt{2\pi}\sigma(x)}\exp\left(-\frac{(\mu(x) - f(x) - EI)^2}{2\sigma^2(x)}\right) \tag{2.21}$$

The Expected Improvement (EI) is the integral over this function:

$$EI = \int_{I=0}^{I=\infty} I \frac{1}{\sqrt{2\pi}\sigma(x)} \exp\left(-\frac{(\mu(x) - f(x) - EI)^2}{2\sigma^2(x)}\right) dI \tag{2.22}$$

As shown by Jones, Schonlau and Welch (1998), this function can be evaluated analytically, yielding:

$$EI(x) = \begin{cases} (\mu(x) - f^*)\Phi(Z) + \sigma(x)\phi(Z) & if\ \sigma(x) > 0 \\ 0 & if\ \sigma(x) = 0 \end{cases} \tag{2.23}$$

$$Z = \frac{\mu(x) - f^*}{\sigma(x)} \tag{2.24}$$

where ϕ and Φ are respectively the probability density function and cumulative density function of the standard normal distribution.

2.2.1.4.Performance metrics and feature importance

Performance of regression models, specially ML-based ones, is commonly measured with some error measures, like R^2, Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE), and Standard Deviation of Absolute Percentage Error (SDAPE), given by:

$$\mathrm{R}^2 = 1 - \frac{\sum_i (\hat{y}_i - y_i)^2}{\sum_i (\bar{y} - y_i)^2} \tag{2.25}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |\hat{y}_i - y_i| \tag{2.26}$$

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \tag{2.27}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2} \tag{2.28}$$

$$\mathrm{MAPE} = \frac{100}{n}\sum_{i=1}^{n} \left|\frac{\hat{y}_i - y_i}{y_i}\right| \tag{2.29}$$

$$\text{SDAPE} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(100\left|\frac{\hat{y}_i - y_i}{y_i}\right| - \text{MAPE}\right)^2} \tag{2.30}$$

where n is the number of samples, y_i is an observed value of the response variable, $\hat{y}_i$ is a predicted value of the response variable and $\bar{y}$ the mean of the observed values.

In this final stage, with the fitted model, feature importance technique may be performed, like Permutation Feature Importance (PFI), in order to answer what features present the biggest impact on predictions of the response variable. Based on Breiman (2001) and Fisher, Rudin, and Dominici (2019), PFI algorithm measures the importance of a feature based on the effect that random disturbances on this feature have on model performance. If performance decreased a lot with the disturbances, the feature is very important, if performance was not affected, the feature is irrelevant for the predictions. Model performance, in regression problems, is usually evaluated with MSE. As shown in Figure 2.7, PFI randomly shuffles the instances of a single feature column and measures the difference between model error before and after the permutation (ΔMSE), the process is repeated N times for robust results based on the mean ΔMSE of the N results. The technique uses the trained model, thus not demanding additional computational resources for retraining.

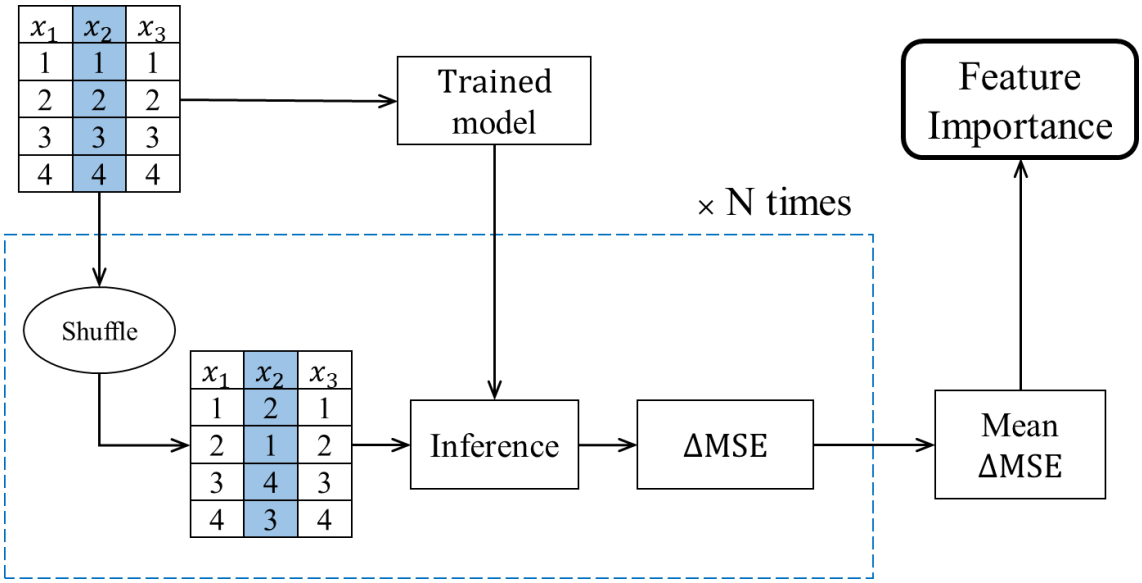


Figure 2.7. Permutation Feature Importance technique

2.2.2. Steel bolted angles net section capacity

Many different equations and models try to predict accurately the resistance of the net section of SBAs connected by one leg, the ones that were used in the current work are presented in this section. Figure 2.8 illustrates a steel angle connected to a gusset plate, showing connected and unconnected legs, location of the critical cross-section, where the holes are located, and some variables that will be shown in equations, like the diameter of the holes, d_h. The load, P, is applied in z axis, thus it being considered the longitudinal direction, as y axis is considered the transverse direction, perpendicular to the load. Figure 2.8 also illustrates the critical cross-section with geometric variables that were used in the present study.

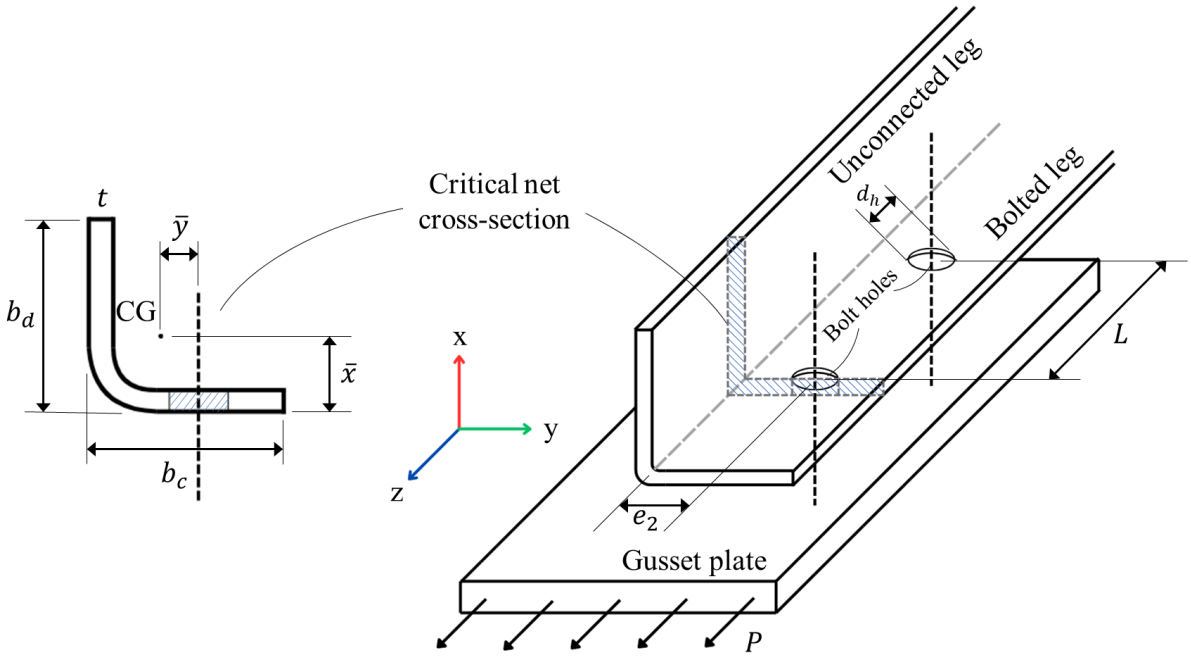


Figure 2.8. Illustration of BSA connected to gusset plate

When available, specific cold-formed steel standards were preferred over general steel structure standards. All equations presented do not include factors of safety, thus providing nominal resistance, not design resistance. The reason for this is one of the objectives of this work, that is to compare predictions of the standards with experimental values, so these predictions should be as accurate as possible, not conservative.

United States of America (USA), Canada and Mexico adopt American Iron and Steel Institute (AISI) standards, using S100-16 Standard (AISI, 2016) for the design of cold-formed steel structural members. The net section capacity of BSAs P_{nt}, that is, the value of the load P when

net section failure occurs, specifically for elements of thickness less than or equal to 4.76 mm connected by one leg is given by

$$P_{nt} = \left(\frac{1}{1.1 + \frac{0.5 b_d}{b_c + b_d} + \frac{2\bar{x}}{L}} \right) A_{nt} F_u \tag{2.31}$$

where b_c is the width of the connected leg, b_d the width of the non-connected leg, L the length of connection in longitudinal direction, A_{nt} the net area of the cross-section, F_u the tensile strength of the steel material and $\bar{x}$ the eccentricity from shear plane to centroid of the cross-section.

In Europe, the European Committee for Standardization (CEN, Comité Européen de Normalisation, in French) established EN 1993-1-8:2005 (CEN, 2005a), that presents three equations for BSA with one line of bolts, for respectively 1, 2 and 3 bolts:

$$P_{nt} = \begin{cases} 2(e_2 - 0.5 d_h) t F_u, & if\ 1\ bolt \\ \beta_2 A_{nt} F_u, & if\ 2\ bolts \\ \beta_3 A_{nt} F_u, & if\ 3\ bolts \end{cases} \tag{2.32}$$

and, in EN 1993-1-1:2005 (CEN, 2005b), one equation for two lines of bolts:

$$P_{nt} = 0.9 A_{nt} F_u \tag{2.33}$$

where t is the thickness of SBA's section, d_h the hole diameter for the bolts, e_2 the distance from center of the hole to the edge of the SBA in transverse direction and β_2 and β_3 are reduction factors, (not factors of safety) dependent on p_1, the longitudinal distance between two consecutive bolts center to center, that follow values in Table 2.1. For $2.5 d_h < p_1 < 5.0 d_h$, β_2 and β_3 may be determined by linear interpolation. Also, for unequal angles connected by its smaller leg, A_{nt} should be taken as the net section area of an equivalent equal-leg angle of size equal to that of the smaller leg.

Table 2.1. Values of reduction factors β_2 and β_3

Pitch p_1	**≤ 2.5 d_0**	**≥ 5.0 d_0**
β_2 (2 bolts)	0.4	0.7
β_3 (3 bolts)	0.5	0.7

Paula *et al.* (2008), based on past experiments reported in the literature (YIP; CHENG, 2000; MAIOLA, 2004), stated that the equation proposed by LaBoube and Yu (1995), given by

$$P_{nt} = \left(1 - 1.20\frac{\bar{x}}{L}\right) A_{nt} F_u \tag{2.34}$$

could be modified for a better representation of the shear lag phenomenon. Also, the authors doubted that the simple equation could describe accurately a complex phenomenon like shear lag, without considering SBA section's thickness, legs' widths or bolt diameter. Motivated by this, they proposed the following equation

$$P_{nt} = \left(1.19 - 0.26\frac{\bar{x}}{L} - \frac{0.63(b_c - n_b d_h) + 0.17 b_d - 0.47 d - 1.7 t}{b_c}\right) A_{nt} F_u \tag{2.35}$$

where n_b is the number of bolts present in the cross-section, that is, the number of bolt lines, d_h, as said before, is the hole diameter, and d is the nominal bolt diameter.

Teh and Gilbert (2013a), based on their own research on net section tension capacity of cold-reduced sheet steel channel braces bolted at the web (TEH; GILBERT, 2013b), adapted their proposed equation for channels to be used on angles, resulting in an equation of similar format to the actual equation presented by AISI Standards and given by

$$P_{nt} = \left(\frac{1}{1.1 + \frac{b_d}{b_c + b_d} + \frac{\bar{x}}{L}}\right) A_{nt} F_u \tag{2.36}$$

The authors stated that the equation takes into account in-plane and out-of-plane shear lag effects, detrimental bending moment effect caused by the connection eccentricity and counteracting bending moment effect. Moreover, they stated that the equation is simples, continuous and transparent, showing no artificial lower or upper bound values for the efficiency

reduction factor that go along $A_{nt}F_u$.

Fleitas *et al.* (2020) performed a parametric study based on finite element models implemented in ABAQUS software, considering geometric nonlinearity, contact between different parts and material nonlinearity. They investigated individually the relation of $\bar{x}$, $\bar{y}$, L and L_t with the efficiency factor of the net section resistance of BSA under tension. The authors verified that there was a linear relationship between them. Additionally, they investigated the relation of the ratios $\bar{x}/L$, $\bar{y}/L$, $\bar{x}/L_t$ and $\bar{y}/L_t$ with the mentioned factor and observed a quasi-linear relation. Based on these results, they performed regression analysis on the studied data and proposed an equation with two sets of coefficients, 0.9, 0.2, 0.1, 0.26 and 0.05 for one bolt lines and 0.9, 0.025, 0.2, 0.3 and 0.15 for two bolt lines, given by

$$P_{nt} = \left(k_0 - k_1\frac{\bar{x}}{L} - k_2\frac{\bar{x}}{L_t} - k_3\frac{\bar{y}}{L} - k_4\frac{\bar{y}}{L_t}\right)A_{nt}F_u \tag{2.37}$$

where L_t is the distance from bolts in transverse direction center to center or, in case of only one bolt, it is equal to the nominal bolt diameter d.

2.2.3. Temperature variation in concrete pavements

Based on heat transfer theory and assuming that the surface could reach an equilibrium state, the calculation of net rate of heat flow to and from a body, q_{net}, can be calculated from

$$q_{net} = q_s + q_a \pm q_c \pm q_k - q_r \tag{2.38}$$

where q_s is the energy absorbed from direct solar radiation, q_a the energy absorbed from diffuse radiation, q_c the energy transferred to or from the body due to convection, q_k energy transferred to or from the body due to conduction, and q_r energy emitted from body through outgoing radiation. The terms q_c and q_k were assumed to be negative because of the general warm climate of Brasília and Brazil in general, thus the pavement surface temperature tends to be higher than at other depths. The result in the following equation

$$q_{net} = q_s + q_a - q_c - q_k - q_r \tag{2.39}$$

Energy absorbed from direct solar radiation, q_s, can be calculated as

$$q_s = \alpha_a R_0 {\tau_a}^{1/\cos z} \cos i$$
$$\cos z = \sin \phi_l \sin \delta_s + \cos \delta_s \cos h \cos \phi_l \qquad 2.40\text{(a-b)}$$

where α_a is the surface absorptivity, R_0 the solar constant, τ_a the transmission coefficient for unit air mass, z the zenith angle, i the angle between the normal to the surface and the direction of radiation, ϕ_l the latitude, δ_s the solar declination, and h the hour angle. In horizontal surfaces, $\cos i = \cos \phi_l$.

Atmospheric radiation absorbed by the pavement surface, q_a, was calculated through an empirical formula developed by Geiger (1959) and reported by Dempsey (1970) given by

$$q_a = \varepsilon_a \sigma_0 {T_{air}}^4$$
$$\varepsilon_a = G - J(10^{-\rho P}) \qquad 2.41\text{(a-b)}$$

where G, J, and P are constant values of respectively 0.77, 0.82, and 0.0074, according to Geiger (1959), σ_0 is the Stefan-Boltzman constant, T_{air} the air temperature, and ρ the vapor pressure varying between 1 and 10 mm of mercury.

The conduction rate of heat flow from the pavement surface down, q_k, was approximately calculated as

$$q_k = -k_c \frac{T_d - T_{surf}}{h_d} \qquad (2.42)$$

where k_c is the thermal conductivity, h_d the depth, T_d the temperature at depth h_d, and T_{surf} the surface temperature. The rate at which the surface emits radiation, q_r, is given by

$$q_r = \varepsilon \sigma {T_{surf}}^4 \qquad (2.43)$$

where ε is the emissivity. The rate of heat flow by convection to the surrounding air, q_c, is given by

$$q_c = h_c(T_{surf} - T_{air}) \qquad (2.44)$$

where h_c is the surface coefficient of heat transfer, which depends on geometry of the surface, wind velocity, and physical properties of the air.

Considering an equilibrium state, the net rate of heat flow, q_{net}, is equal to zero. So, getting all the mentioned equations together, the equilibrium temperature can be obtained by

$$R_0 \alpha_a \tau_a^{\frac{1}{\cos z}} \cos z + \varepsilon_a \sigma_0 {T_{air}}^4 - h_c\left(T_{surf} - T_{air}\right) - \frac{\mathrm{k_c}}{h_d}\left(T_{surf} - T_d\right) - \varepsilon \sigma_0 {T_{surf}}^4 = 0$$

$$\cos z = \sin \phi_l \sin \delta_s + \cos \delta_s \cos h \cos \phi_l \qquad 2.45\text{(a-c)}$$

$$h_c = 698.24\left[0.00144\mathrm{T_m^{0.3}U^{0.7}} + 0.00097\left(T_{surf} - T_{air}\right)^{0.3}\right]$$

where T_m is the average of the surface temperature T_{surf} and air temperature T_{air} in ºK, and U the average daily wind velocity in m/s. Equation 2.45(c) is an empirical formula developed by Vehrencamp (1953) and reported by Dempsey (1970).

3. METHODOLOGY

The present chapter describes all of the methodology related to Machine Learning modeling done in this thesis. First, it explains baseline models and, then, explains the implemented SOFAH algorithm.

All programming activities were executed by the author of the present document in Python 3.12.0, developing in-house code with assist of some libraries:

- Matplotlib and seaborn for plotting and data visualization;
- Pandas and NumPy for data manipulation and analysis
- Scikit-learn for inputs normalization, training-validation-test splitting and performance measurement;
- Scikit-optimize for hyperparameters tuning and optimization;
- Joblib for parallelization;
- Finally, PyTorch for all activities related to ANN modeling, some of them are weights initialization, training, validation, testing, construction of neural network architecture, implementation of regularization techniques, etc.

The general methodology adopted for each engineering problem analyzed is very similar, Figure 3.1 shows an example using the Steel Bolted Angles (SBA) problem. Initially, the importance of the engineering problem is evaluated, as, in the example, SBAs are commonly connected by only one leg in transmission and communication towers. Then, a dataset is gathered, being it composed of numerical or experimental data. An ANN model is developed based on the collected data, resulting in a predictive model. The performance of these models and the importance of each feature used are measured, providing error quantification and information on most relevant features for prediction of the output.

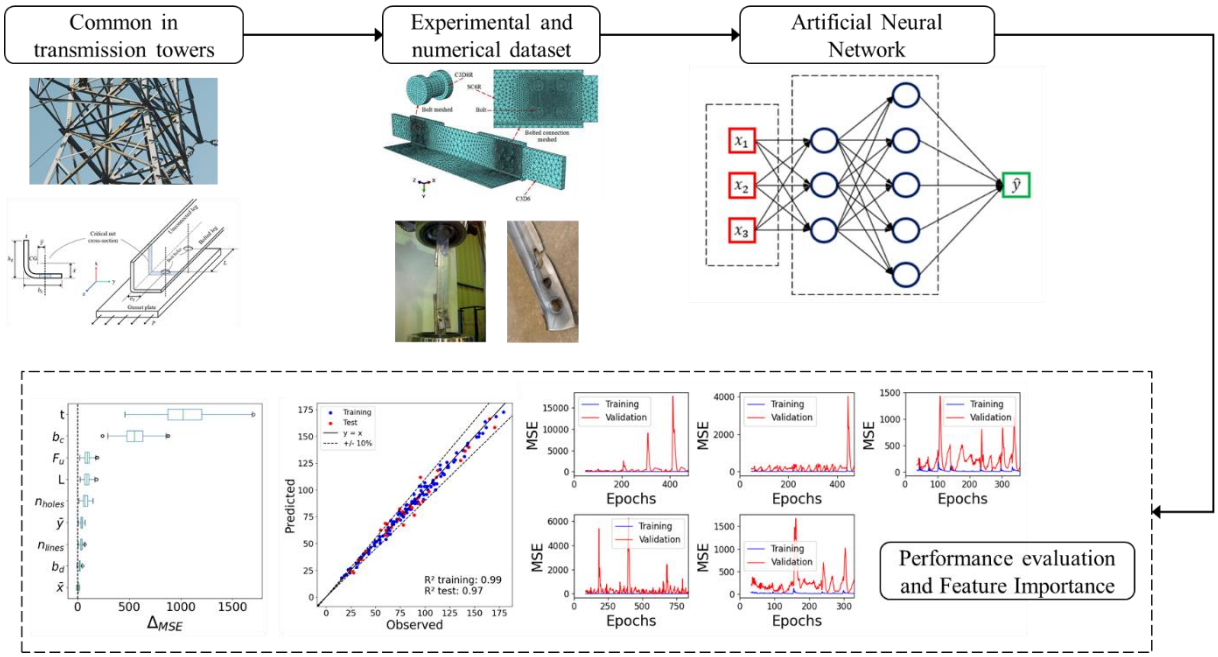


Figure 3.1. General steps of adopted methodology

3.1. BASELINE MODELS

In this section, many specific details of the baseline methodology of ML modeling are explained. The general steps are shown in Figure 3.2, where initially a dataset is gathered, data exploration is performed and the possibility of feature selection is investigated. These data-related steps are iterative, as mentioned in section 2.2.1, because of the impact of high-quality and abundant data in ML modeling. After the establishment of a good database, model training stage is initiated, starting with the definition of the loss function, the optimizer for model's parameters and method of parameter initialization. Then, cross-validation is used in hyperparameters optimization phase, while the model is trained in multiple folds (k-fold cross-validation was used). Regularization techniques are fundamental to avoid overfitting in model training. With the trained model, is final performance is measured on unseen instances, that is, on the test set. This performance analyses always come accompanied of graphical and residuals analyses, while also evaluating feature importance. This described methodology, in its development phase can be seen as iterative, because, if accuracy is not satisfactory, many variables can be modified and the entire methodology repeated to provide a better final model.

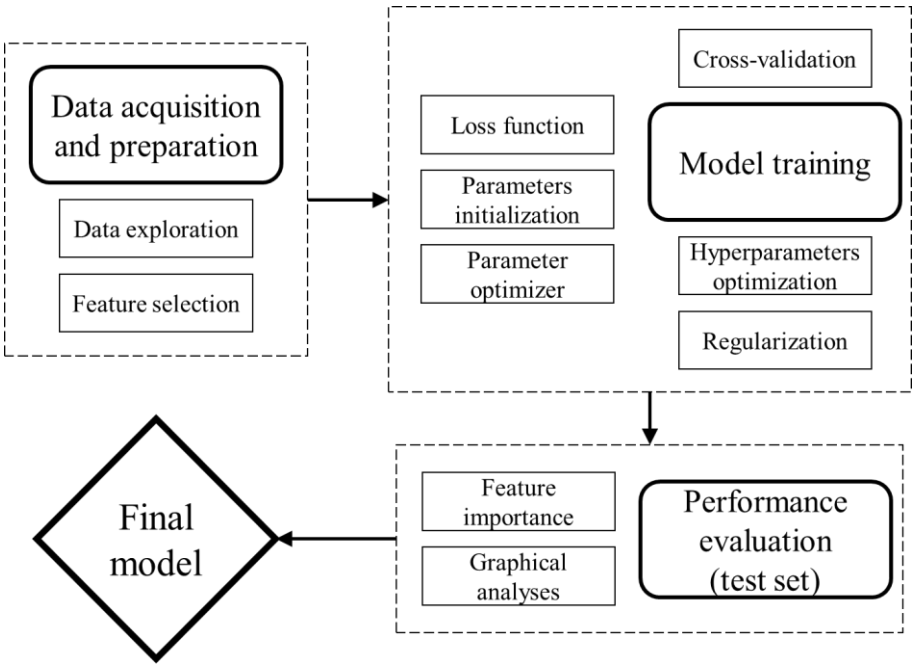


Figure 3.2. General steps of the baseline methodology for ML modeling

The data was fed through a feedforward fully connected Artificial Neural Network (ANN), also known as a Multilayer Perceptron (MLP) architecture. The input layer and every hidden layer of the MLP were built sequentially with a linear transformation (applying weights and biases), a batch normalization layer and, a ReLU (Rectified Linear Unit) activation function. In reason of this choice of activation function, initialization of linear layers weights was done with Kaiming He Initialization, while biases were initialized to zero. In batch normalization layers, weights were initialized to ones and biases to zeros. Training was performed with AMSGrad variant of the Adam optimizer until convergence.

The criteria of convergence were both maximum number of epochs and patience number of epochs in early stopping technique. For most of the datasets used, bolted steel angles, yacht hydrodynamics and energy efficiency, all models defined 10,000 as the maximum number of epochs and 200 as patience. For concrete pavement temperature, since it is a much larger dataset, maximum of epochs was established as 300 in HPO and 1000 in training, and patience as 20. These high values of maximum epochs were used to guarantee convergence during training, since the implementation of early stopping prevented overfitting and stopped training much before the established maximum of epochs.

In concrete pavement temperature dataset, the mentioned choice of smaller values of epochs was due to batch size. In their models, batch size of 256 and 512 were used and, since they are composed of thousands of instances, a large number of weight updates was performed in each epoch. This means that, in these models, a smaller number of epochs was needed for convergence. In the rest of the datasets, because of their small number of instances, batch size was set as the number of instances in the training set of cross-validation. Since all models used 5-fold cross-validation, this number is equal to 80% of the whole training set size.

Bayesian Optimization (BO) was used to optimize hyperparameters listed in Table 3.1 with their respective search space in all developed models. The technique was implemented with parallelization, where each iteration of the algorithm used 2 CPU cores, each with 2 processes, to evaluate 4 different points of the search space simultaneously. The 12 first points, that is, the 3 first iterations, are chosen randomly, and only after that the objective function is approximated with Gaussian Processes and new points are chosen based on the acquisition function Expected Improvement (EI). For this parallelization of BO algorithm, a multi-point EI with constant liar min strategy was implemented (CHEVALIER and GINSBOURGER, 2012).

The algorithm was programmed to perform 50 iterations, that is, evaluate 200 points, or stop early if the criterion of convergence was met, no new point of minimum in 50 consecutive points evaluated. The objective function used was the mean of the Mean Squared Errors (MSE) of the 5 validation folds.

Table 3.1. Optimized hyperparameters and respective search space

Hyperparameter	**Range**
Learning rate (α)	$[10^{-4}, 10^{-1}]$
Weight decay (λ)	$[10^{-9}, 1]$
Hidden layers	[1, 4]
Neurons in each hidden layer	[10, 1000]

Parallelization was also implemented for model training and inference on GPU with PyTorch's support to CUDA, a parallel computing platform and programming model created by NVIDIA that allows for straightforward GPU-accelerated data-driven modeling. All models of the smaller datasets were trained on present author's personal computer equipped with a NVIDIA GeForce MX110 of only 2 GB of dedicated memory. Models of the pavement temperature

dataset were trained on a high-performance computer of the Laboratory of Computational Methods and Artificial Intelligence located at the Postgraduate Program on Structures and Civil Construction at the University of Brasília, equipped with a NVIDIA GeForce RTX 4060 Ti of 16 GB of dedicated memory.

When performing k-fold cross-validation with early stop, each fold will have its training stopped at a different number of epochs, so, after performing hyperparameters optimization, if the whole training set is being used and there is no validation set, it is not intuitive at how many epochs training should be stopped. Motivated by this problem, all models developed in this work act like an ensemble of 5 models, using the training and validation sets of each fold of the cross-validation. So that, training is stopped based on each validation set and inference is performed calculating the mean of predictions given by each model.

3.2. SIMULTANEOUS OPTIMIZATION OF FEATURE AUGMENTATION AND HYPERPARAMETERS (SOFAH)

Simultaneous Optimization of Feature Augmentation and Hyperparameters (SOFAH) is an algorithm proposed by the author of the current thesis that addresses both Feature Selection (FS) and Hyperparameter Optimization (HPO) problems concurrently. SOFAH aims to reliably and automatically select either the optimal subset of features or a subset of very similar performance. As many Feature Selection algorithms are exhaustive, performing a large amount of evaluations in search of the optimal subset configuration, SOFAH uses BO, an optimization technique known for being good at evaluating expensive black box functions in carefully selected points (NGUYEN, 2019).

There is a lack of studies that apply BO in FS problems, although a very recent one showed promising results (YANG; LIU; WEN, 2024). There are indications in the literature that BO is best-suited for optimization over continuous domains of at most 20 dimensions (FRAZIER, 2018; NGUYEN, 2019), but, also, there is a very recent published research that ventured on high-dimensional spaces and obtained good results (GÖRMEZ, 2024). Furthermore, this problem of the high dimensionality was ignored because SOFAH algorithm do not aim to necessary obtain an optimal solution, as a close-to-optimal solution suffices.

The idea of using Bayesian Optimization to perform Feature Selection is not new. However, performing feature generation to augment the feature space and then applying BO to concurrently address HPO and FS could not be found in any other published research until the present moment.

The steps of the algorithm consist in initially performing feature generation, this is done with square, cube, inverse and logarithm transforms and also performing multiplication between two variables (Figure 3.3). Features that present zeros are excluded from inverse and logarithm transforms. After this step, an augmented feature space of added information was formed. Then, when optimizing the hyperparameters described in the previous section, 3.1, a large number of dimensions is added to the search space, being the first one the number of features and each one of the others a feature. This means that the algorithm finds the optimal number of features and also which features these will be.

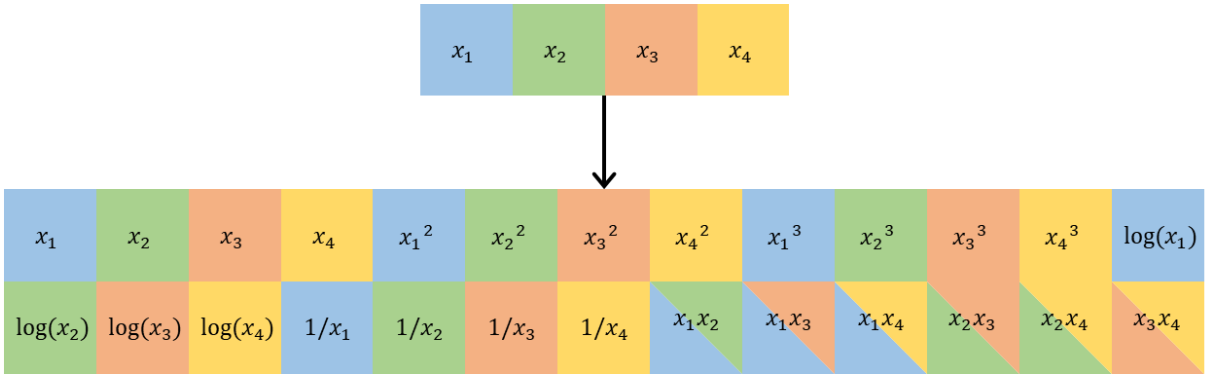


Figure 3.3. Illustration of feature augmentation in SOFAH algorithm

Observing Figure 3.2, shown in the previous section, 3.1, it can be said that SOFAH is a simple algorithm, since it only increases the dimensions of the feature space and of the search space in HPO phase, that has added dimensions to search for optimal number of inputs and optimal inputs. All the other steps of ML modeling baseline methodology are kept unchanged in the proposed algorithm. Still regarding to the mentioned figure, FS and HPO steps can be seen in the baseline methodology steps. FS step consisted of performing manual feature engineering based on data exploration results, while in SOFAH, FS is an automated process based on a powerful optimization technique.

Let the hyperparameters search space S_{HP} be

$$S_{HP} = \{\propto, \lambda, n_{layers}, n_n^1, n_n^2, n_n^3, n_n^4\} \quad (3.1)$$

where n_{layers} is the number of hidden layers, and n_n^1 to n_n^4 are the number of neurons in each layer. Now, let the search space S_{orig} of original features $x_1, x_2, x_3, \dots, x_n$ be

$$S_{orig} = \{x_1, x_2, x_3, \dots, x_n\} \quad (3.2)$$

The augmented feature search space S_{FS} can be defined as a union between S_{orig} and every set of transformation performed on the original features, that is,

$$S_{FS} = S_{orig} \cup S^2 \cup S^3 \cup S^{-1} \cup S^{log} \cup S^{mul} \quad (3.3)$$

where S^2 is the set of square transformation, S^3 the set of cube transformation, S^{-1} the set of inverses, S^{log} the set of logarithm transformation, and S^{mul} the set of multiplications between two original variables. Finally, the search space of SOFAH methodology S_{search} is defined as

$$S_{search} = S_{FS} \cup S_{HP} \quad (3.4)$$

Representing the SOFAH algorithm in a pseudocode, in Figure 3.4, it can be seen that it is very similar to the presented pseudocode for BO, the main difference is the mentioned implementation of parallelization and the different search space. As mentioned, the methodology was implemented with $\text{N\_ITERS} = 50$ iterations, each evaluating $\text{N\_POINTS} = 4$ points in parallel and with $\text{PATIENCE} = 50$. Also, the $random$ function now samples randomly multiple points, because of the parallelization. As stated, the main difference from baseline methodology is the search space composed of many dimensions related to optimal features. Also, the vector of initial points $\boldsymbol{x_0}$ include these extra dimensions. The exponent *j* represents an instance of $\boldsymbol{x_i}$ or $\boldsymbol{y_i}$ vectors.

inputs: S_{search}, $\boldsymbol{x_0}$, N_ITERS, PATIENCE, N_POINTS

functions: $objective(x_i^j)$, $random(S, n)$, $acq\_func(x)$, $GP(x, y)$

LOSS_MIN $= \infty$

$x, y = list(), list()$

for $i = 0:$ N_ITERS $- 1:$

    for $j = 1:$ N_POINTS:

        $y_i^j = objective(x_i^j)$

        $x, y = x.append(x_i^j), y.append(y_i^j)$

    if $\min(y_i) <$ LOSS_MIN:

        LOSS_MIN $= y_i^j$, where j is the point where $y_i^j = \min(y_i)$

        $x\_min = x_i^j$, where j is the point where $y_i^j = \min(y_i)$

        TOL $=$ PATIENCE

    else:

        TOL $=$ TOL $-$ N_POINTS

        if TOL $= 0$:

            break

    if $i < 10$:

        $x_{i+1} = random(S_{search}, \text{N\_POINTS})$

    else:

        $GP(x, y)$

        $x_{i+1} = \text{argmax}\, acq\_func(x)$

return x_min

Figure 3.4. Pseudocode of SOFAH algorithm

Note that S_{FS} has many more dimensions than S_{HP}, so, to counter a possible loss in accuracy due to very high-dimensional search space in HPO, another version of the SOFAH algorithm was proposed, named SOFAH2. This second version adds a second step to standard SOFAH, where an additional HPO phase is performed but with none of the added dimensions related to optimal feature selection. This two-step version of the algorithm allows for finer tuning of the hyperparameters, performing a second search but now in low-dimensional space. It is important

to point that in this second step, the inputs being used are the optimal ones, found by the optimization of step one. Also, the hyperparameters found as optimal in step one are used as an initial point for BO in step two.

The added amount of computational resources used in SOFAH2, that requires first executing SOFAH, should achieve greater accuracy than SOFAH to be viable. But, in general, the computational cost of SOFAH and SOFAH2 is highly dependent on the number of iterations performed by the BO until convergence. It is not possible to affirm beforehand if it takes more time than the baseline methodology.

In the present work, in every use of SOFAH and SOFAH2, the optimal number of features was searched from 1 to the total number of features (after augmentation). Furthermore, one of the initial points was always, in part, chosen deterministically, as the hyperparameters were still random but features-related dimensions were fixed. The number of optimal features was always initiated as half of the total number of features and the optimal inputs initiated as the most correlated features with the output. Finally, all other phases of the methodology for ML modeling were done exactly the same as described in the baseline methodology, in previous section, 3.1.

4. UC IRVINE REPOSITORY DATASETS

For further comparison between baseline model methodology and Simultaneous Optimization of Feature Augmentation and Hyperparameters (SOFAH) algorithm, some publicly available datasets were used as real-world benchmarks. These datasets can be downloaded for free at the UC (University of California) Irvine Machine Learning Repository site (KELLY; LONGJOHN; NOTTINGHAM, n.d.). An advantage of using these public datasets is that other researchers already implemented a wide range of algorithm on them, thus providing results to be compared with the ones obtained in the present research.

4.1. YACHT HYDRODYNAMICS

The Yacht Hydrodynamics (YAH) dataset (GERRITSMA; ONNINK; VERSLUIS, 1981) consists of 308 instances of results of full-scale experiments performed on ships with 22 different hull forms at the Delft Ship Hydromechanics Laboratory. It presents 7 variables, all dimensionless, the inputs concern the Froude number and 5 hull geometry coefficients, and the output the of residuary resistance per unit weight of displacement, as shown in Table 4.1.

Table 4.1. Variables in yacht hydrodynamics dataset

Variable	Description
C_P	Longitudinal position of the center of buoyancy
P_C	Prismatic coefficient
L/d	Length-displacement ratio
B/d	Beam-draught ratio
L/b	Length-beam ratio
F	Froude number
R_R	Residuary resistance per unit weight of displacement

The residuary resistance, R_R, of sailing yachts at the initial design stage is very valuable for evaluating ship's performance and for estimating the required propulsive power. As affirmed by Rawson and Tupper (2001), it consists of wave-making resistance, form resistance, eddy resistance and frictional form resistance, and is given by

$$R_R = R_T - R_F \tag{4.1}$$

where R_T is the total resistance and R_F the frictional resistance of an equivalent flat plate.

Three ML-based models were trained for YAH dataset, using all the 7 variables and, respectively, with baseline methodology, single-step SOFAH algorithm and two-step SOFAH (SOFAH2). Training-test split remained the same as in the other chapters of the current study 80% for training (246 instances) and 20% for test (62 instances). Some regression evaluation metrics for training and test sets are shown in Table 4.2, Mean Absolute Error (MAE), RMSE, R^2, MAPE and Standard Deviation of Absolute Percentage Error (SDAPE).

Table 4.2. Regression evaluation metrics of YAH dataset

Performance metrics	$\mathbf{M_{YAH}}$	$\mathbf{M_{YAH}^{SOFAH}}$	$\mathbf{M_{YAH}^{SOFAH2}}$
MAE - Training	0.120	0.118	0.135
MAE - Test	0.211	0.267	0.231
RMSE - Training	0.182	0.179	0.174
RMSE - Test	0.533	0.540	0.479
R^2 - Training	1.000	1.000	1.000
R^2 - Test	0.998	0.998	0.998
MAPE - Training	23.97%	9.37%	33.82%
SDAPE - Training	163.94%	49.92%	205.37%
MAPE - Test	39.26%	28.38%	54.25%
SDAPE - Test	127.77%	83.45%	164.18%

As observed, all the developed models presented outstanding performance, with R^2 of 0.998 on test set. Following the successive order of M_{YAH}, M_{YAH}^{SOFAH} and M_{YAH}^{SOFAH2}, RMSE always went down in training set, as in test set it went up in M_{YAH}^{SOFAH} but down in M_{YAH}^{SOFAH2}. R^2 stayed the same, but only because of its already extremely good value close to its upper bound of 1.000. MAPE metric presented bad results, but SDAPE is also very high, indicating that there is high dispersion and possibly just a few points of high APE. Despite being widely used, MAPE is biased towards low predictions and cannot be used with zero or close-to-zero values. Based on exploratory analysis, the minimum value of the response variable is 0.01, its first quartile is 0.78 and its skewness is 1.76, thus presenting a considerable amount of close-to-zero values. MAE is only shown for comparison purposes, with results from other authors.

From these results, single-step SOFAH algorithm did not improve error metric RMSE when compared to M_{YAH}, however, SOFAH2 was able to improve the already excellent model M_{YAH},

proving that the algorithm is capable of achieving high performance in regression problems.

Predictions of model M_{YAH}^{SOFAH2} are shown in Figure 4.1 and Figure 4.2, compared to observed values, residuals and relative residuals. In Figure 4.1, it is clear that the variable is very right skewed and that there is an almost perfect fit to both training and test sets. Based on Figure 4.2 a), there is only point that, compared to the others, show high residual. Based on Figure 4.2 b), outliers of very high positive relative residuals can be seen all in close-to-zero values, reaffirming the reason for bad MAPE metrics in the highly accurate developed models.

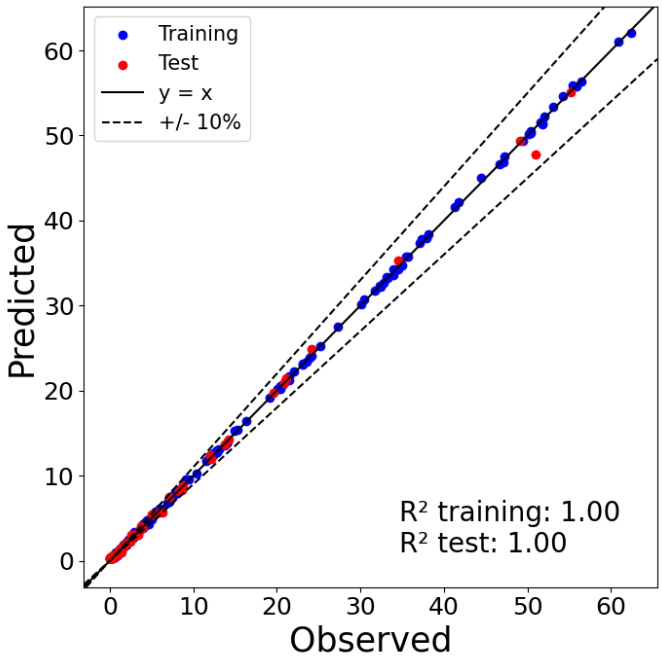


Figure 4.1. Predicted vs. observed plot for training and test for model M_{YAH}^{SOFAH2}

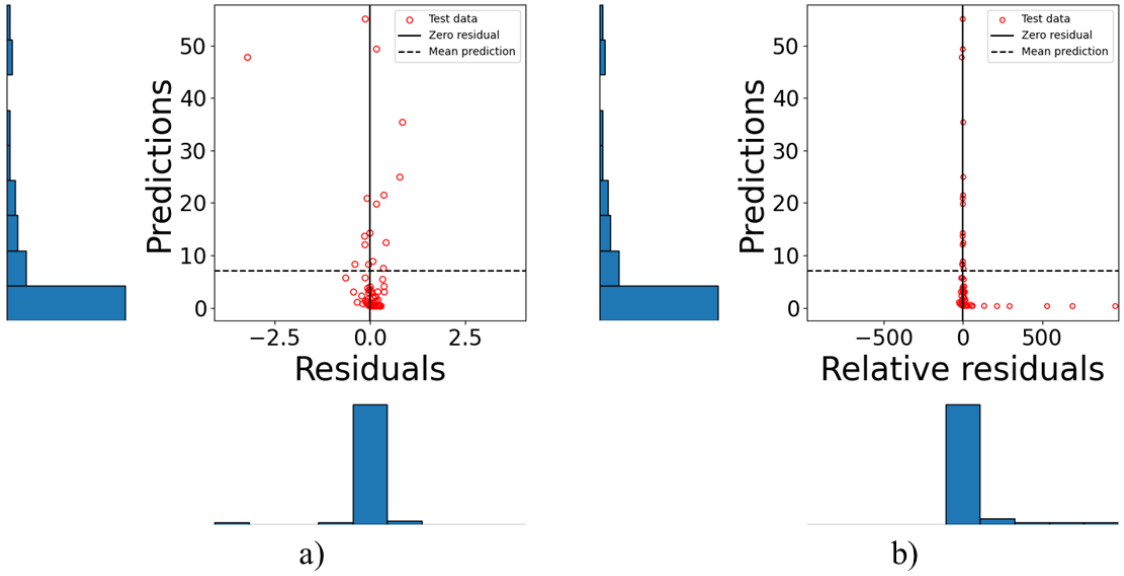


Figure 4.2. Model M_{YAH}^{SOFAH2} on test set predictions and: a) Residuals; b) Relative residuals

Fan *et al.* (2019) presented a Two-Layer Wang-Mendel (TL-WM) fuzzy approach to improve the traditional Wang-Mendel (WM) model (WANG L. X., 2003). The WM model is a widely used fuzzy system due to its simplicity, high interpretability and efficiency in extracting fuzzy rules from numerical data (FAN *et al.*, 2019). Fan *et al.* (2019) used the YAH dataset to compare performance of its proposed TL-WM model with eight other algorithms: three well known ML techniques, Random Forest, Artificial Neural Network and Support Vector Machine; traditional WM model; improved WM based on Fuzzy C-Means (GOU *et al.*, 2015); improved WM based on Fast Search and Find of Density Peaks clustering algorithm and sample correlation (GOU *et al*, 2016); reduced weighted WM (FAN *et al.*, 2016); and fuzzy model identification based on fuzzy-rule clustering (FAN *et al.*, 2017).

Mukhoti *et al.* (2018) published a study on the importance of strong baselines, exposing that the common experimental procedure in ML of benchmarking proposed models against baselines is flawed when the identical experimental setups of proposed models are not used in the baselines. The authors use many UCI datasets (KELLY; LONGJOHN; NOTTINGHAM, n.d.) to compare results from: 3 settings of Neural Networks with Monte Carlo dropout (GAL; GHAHRAMANI, 2016); Bayesian neural networks using Variational Matrix Gaussian posteriors (LOUIZOS; WELLING, 2016); Bayesian networks with horseshoe priors (GHOSH; DOSHI-VELEZ, 2017); and probabilistic backpropagation with the Matrix Variate Gaussian

distribution (SUN; CHEN; CARIN, 2017).

Between all the 9 mentioned models from Fan *et al.* (2019), the best performing one was TL-WM, showing MAE of 1.363 and RMSE of 2.783. The second-best MAE value was of 1.931 and second best RMSE was 3.520. Regarding all 6 models from Mukhoti *et al.* (2018), the best performance was shown in one of the settings of Neural Network with Monte Carlo dropout, where it was achieved a RMSE of 0.67 +/- 0.05, and the second best was another one of these settings, achieving RMSE of 0.70 +/- 0.05. These values are all higher than the ones obtained by the 3 models developed in the current research, M_{YAH}, M_{YAH}^{SOFAH} and M_{YAH}^{SOFAH2}, as a graphical comparison between them is shown in Figure 4.3.

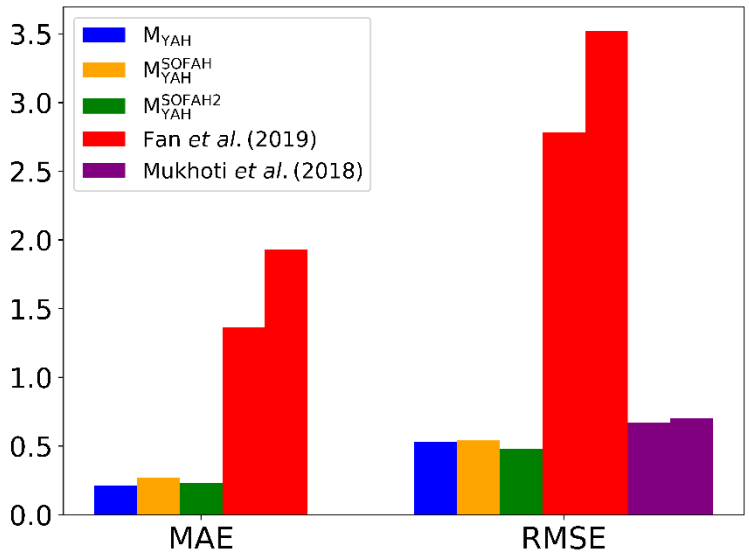


Figure 4.3. Comparison with results of other authors in YAH dataset

4.2. ENERGY EFFICIENCY

The energy efficiency (ENEFF) dataset (TSANAS; XIFARA, 2012a) consists of 12 distinct building shapes simulated in Autodesk Ecotect Analysis, a sustainable design software analysis software that was discontinued by Autodesk, not allowing the acquisition of new licenses since March 2015 due to the integration of energy efficiency and high-performance design tools into the known Revit product family (AUTODESK SUPPORT, 2023). ENEFF general theme is the

design of energy-efficient buildings with improved energy conservation properties (TSANAS; XIFARA, 2012b).

According to the Building Technologies Program of the United States Department of Energy (BURDICK, 2011), HL and CL are the measure of energy needed to be added or removed from a space by the heating, ventilation and air conditioning system in order to achieve the desired level of comfort. The value of these two variables dictate equipment selection and duct design, while also influencing costs, operating energy efficiency, occupant comfort, indoor air quality and building durability.

The dataset is composed of 768 instances and 10 variables, being 8 intended to be used as inputs to predict 2 outputs, Heating Load (HL) and Cooling Load (CL). The description of all variables is shown in Table 4.3.

Table 4.3. Variables in energy efficiency dataset

Variable	Description
R_C	Relative compactness
S_A	Surface area
W_A	Wall area
R_A	Roof area
O_H	Overall height
G_A	Glazing area
O	Orientation
G_A	Glazing area distribution
HL	Heating load
CL	Cooling load

All the developed models for ENEFF used all the 8 inputs available and 1 output. A total of 6 models were trained, 3 for each one of the 2 outputs, HL or CL, being one with baseline methodology and the other two with SOFAH and SOFAH2 respectively. Performance evaluation metrics are shown in Table 4.4, for models that predict HL, and in Table 4.5, for models that predict CL.

Table 4.4. Regression evaluation metrics for HL of ENEFF dataset

Performance metrics	$\mathbf{M_{HL}}$	$\mathbf{M_{HL}^{SOFAH}}$	$\mathbf{M_{HL}^{SOFAH2}}$
RMSE - Training	0.328	0.264	0.264
RMSE - Test	0.454	0.433	0.433
R^2 - Training	0.999	0.999	0.999
R^2 - Test	0.998	0.998	0.998
MAPE - Training	1.18%	0.91%	0.91%
SDAPE - Training	1.20%	0.98%	0.98%
MAPE - Test	1.61%	1.50%	1.50%
SDAPE – Test	1.38%	1.34%	1.34%

Table 4.5. Regression evaluation metrics for CL of ENEFF dataset

Performance metrics	$\mathbf{M_{CL}}$	$\mathbf{M_{CL}^{SOFAH}}$	$\mathbf{M_{CL}^{SOFAH2}}$
RMSE - Training	0.273	0.270	0.216
RMSE - Test	0.787	0.743	0.748
R^2 - Training	0.999	0.999	0.999
R^2 - Test	0.993	0.994	0.994
MAPE - Training	0.82%	0.95%	0.72%
SDAPE - Training	0.76%	0.91%	0.73%
MAPE - Test	1.96%	1.96%	1.90%
SDAPE – Test	1.83%	1.73%	1.71%

As observed, all the developed models presented outstanding performance, with R^2 greater than 0.990 on test set. Following the successive order of baseline model, SOFAH-based model and SOFAH2-based model, RMSE always went down or stayed the same in training and test set for HL, and varied for CL, achieving minimum in M_{CL}^{SOFAH}. Models M_{HL}^{SOFAH} and M_{HL}^{SOFAH2} showed the same metrics because their set of optimal hyperparameters were the same. R^2 stayed the same or increased in test set. MAPE metric and SDAPE always went down, indicating lower and less dispersed percentage errors.

From these results, single-step and two-step SOFAH algorithm improved all error metrics when compared to baselines M_{HL} and M_{CL}, proving again that the algorithm is capable of achieving high performance in regression problems, with even better metrics than the baseline methodology.

Predictions of models M_{HL}^{SOFAH} and M_{CL}^{SOFAH} are shown in Figure 4.4, compared to observed values on training and test sets and with margins of error of +/- 10%. It is clear that the models showed an almost perfect fit to both training and test sets, with only a few points of noticeable prediction error.

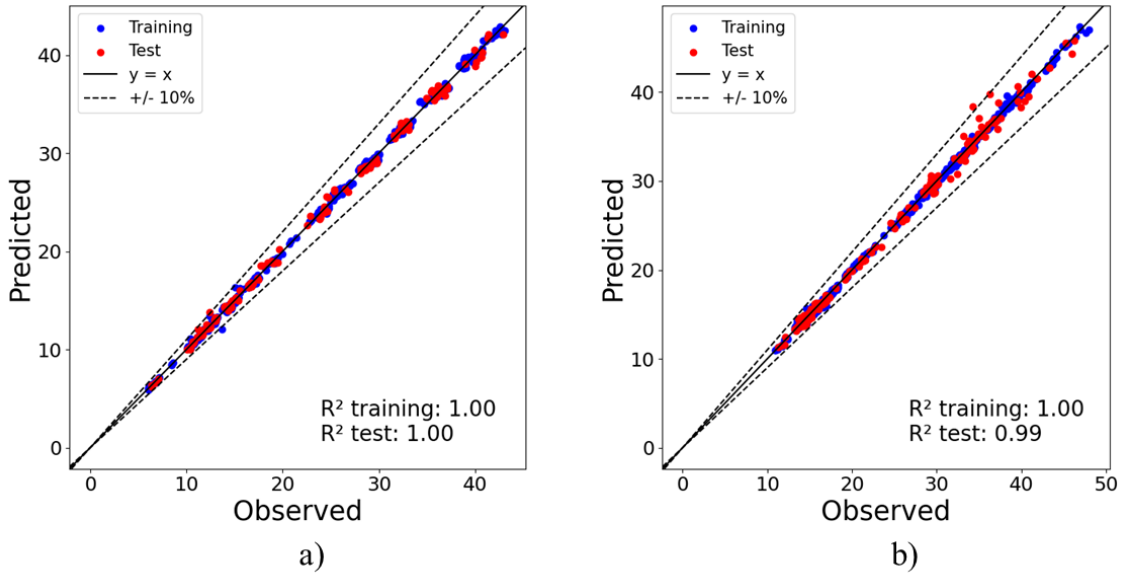


Figure 4.4. Predicted vs. observed values of training and test for: a) M_{HL}^{SOFAH}; b) M_{CL}^{SOFAH}

Žegklitz and Pošík (2017) performed a systematic comparison between symbolic regression and machine learning algorithms on common sets of problems, including ENEFF. The authors, like the present thesis, analyzed separately HL and CL outputs. The investigated algorithms were: Linear Regression, Support Vector Machine Regression, Random Forest, Fast Function Extraction (MCCONAGHY, 2011), Evolutionary Feature Synthesis (ARNALDO; O'REILLY; VEERAMACHANENI, 2015), and Multi-Gene Genetic Programming (HINCHLIFFE *et al.*, 1996).

The authors found that, among all these algorithms Support Vector Machine Regression

provided the best RMSE value for CL, 1.278, and RF the best RMSE for HL, 0.510. The second-best value of RMSE for CL was 1.633 and for HL was 0.546. In the previous section, 4.1, the study of Mukhoti *et al.* (2018) was already mentioned and, as it used a lot of datasets, they used ENEFF, but only for predictions on HL. The best performing analyzed model achieved RMSE of 0.45 +/- 0.01, and the second-best RMSE of 0.54 +/- 0.02. These values are all higher than the ones obtained by the 6 models developed in the current research, as can be seen in Figure 4.5. The CL models showed a considerable difference, but the HL models presented very similar performance with the developed ML models in the present work.

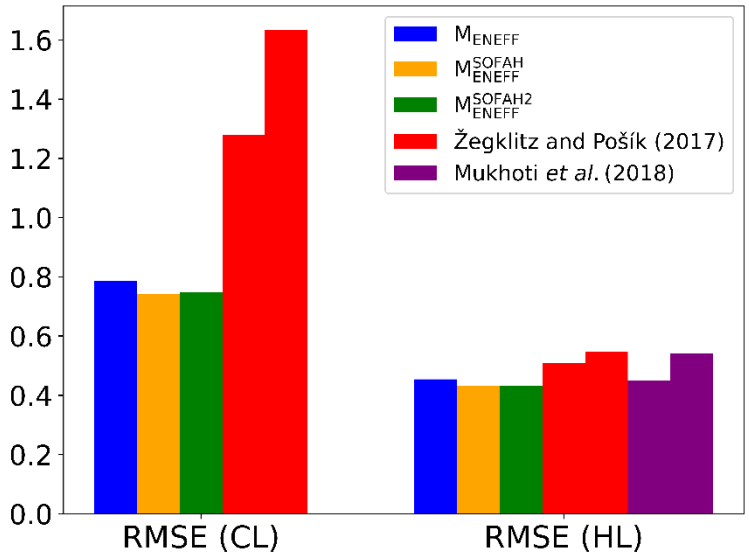


Figure 4.5. Comparison with results of other authors in ENEFF dataset

5. COLD-FORMED STEEL BOLTED ANGLE CONNECTIONS UNDER AXIAL TENSION

The manufacture of iron by ancient civilizations began thousands of years before Christ, used by civilizations in Egypt. However, the use of iron on an industrial scale only occurred in the 19th century, with the advent of the first industrial revolution, especially due to the need of buildings with large spans, warehouses, bridges and railways (PAULA, J. H. M., 2021).

Primarily consisting of iron and carbon, steel is one of the main structural materials in civil engineering. The material shows many advantages over reinforced concrete, such as low structural weight, high precision and technological quality control, reduced construction site, large spans, flexibility for adaptations and possibility of different shapes, generating greater freedom in architectural projects (FONTES, 2020).

Steel, for structural use, can take on different forms, such as sheets, bars and rolled profiles, all manufactured in rolling mills, devices that shape the steel according to the desired cross-section (PFEIL; PFEIL, 2019). A type of steel is cold-formed steel, a material manufactured by bending, rolling or pressing steel, achieving light weight, high strength and ease of installation.

Steel angles are L-shaped laminated steel elements composed of two legs forming a 90-degree angle between them and are commonly used to connect metal parts. The most common ways to connect metal parts to each other are using bolts or welding, both of which have their advantages and disadvantages. Drilling holes for using bolts is expensive and reduces the cross-section of the element. On the other hand, using welding does not allow for the assembly and disassembly of structures, losing versatility and flexibility. (PAULA, J. H. M., 2021).

As mentioned, to use bolted connections, it is necessary to drill holes in the elements, thus the concept of net cross-sectional area arises, which represents the gross cross-sectional area minus the area of the holes. As stated by Fleitas *et al.* (2020), bolted connections in Steel Bolted Angles (SBA) under tension present a complex stress state, therefore the resistance of the net cross-section is not calculated in such a simple way. When bolted and subjected to axial tension, the elements may present net section failure, as one of some possible failure modes, caused by rupture of the section where the holes are located. These elements are widely used in various structures, like, power transmission towers, communication towers, and bridge diagonal

members (QU; GUO; SUN, 2022; LI *et al*., 2024).

SBA members with eccentric connections are common in civil engineering, particularly in lattice towers and masts (QU; GUO; SUN, 2022), thus being influenced by complex debilitating phenomena such as shear lag. The phenomenon causes a nonuniform distribution of the tension stresses across the net section and, added to the effects of eccentricity, acts reducing angles tension capacity. Because of that, in steel structures standards across the world, a reduction factor is applied to the resistance of the net section, while many researches have been conducted over the years trying to obtain better formulations for this factor.

5.1. METHODOLOGY

The current section describes the gathered dataset and both sets of developed models, model set I and model set II. As already mentioned, all models were built with the baseline methodology detailed in section 3.1, unless explicitly specified that it used SOFAH algorithm.

5.1.1. Dataset

The gathered dataset is composed of 314 instances and was collected from five studies, four of which were experimental, composing 154 instances, and one numerical with 160 instances from finite element models implemented in ABAQUS software. The size of each collected sample, type of each study and corresponding authorship are listed in Table 5.1. The dataset composes a 314x10 matrix, each instance presents ten variables (Table 5.2), nine of which were initially used as inputs to predict the target variable T_{obs}.

Table 5.1. Source and number of instances of sampled steel angles data

Authors	Type	Sample size
LaBoube and Yu (1995)	Experimental	12
Yip and Cheng (2000)	Experimental	15
Paula *et al.* (2008)	Experimental	66
Teh and Gilbert (2013)	Experimental	61
Fleitas *et al.* (2020)	Numerical	160
AISI (2016)	Equation (2.31)	-
Eurocode (CEN, 2005a, 2005b)	Equations (2.32)(2.33)	-
	TOTAL	314

Table 5.2. Identification and description of each used feature

Feature	Unit	Description
b_c	mm	Width of the connected leg
b_d	mm	Width of the not connected leg
t	mm	Thickness
n_{lines}	-	Number of bolt lines (1 or 2)
n_{holes}	-	Number of holes per bolt line (1 to 4)
$\bar{x}$	mm	Connection eccentricity, from shear plane to centroid of cross-section
L	mm	Length of connection in longitudinal direction
$\bar{y}$	mm	Connection eccentricity, from centroid of the connection to centroid of cross-section
F_u	MPa	Ultimate tensile strength of the steel material
T_{obs}	kN	Resistance of the net section

In order to stabilize and improve model training, min-max normalization was applied to each input feature, transforming minimum value to 0 and maximum to 1. The transformation parameters were calculated only in the training set of each fold during cross-validation, to guarantee that no data leakage would occur.

It is worth mentioning that not all of the variables used were explicitly available in the mentioned studies, requiring some calculations. To calculate $\bar{y}$, it was assumed that the bolt connection was always exactly in the middle of the connected leg, that is, in $b_c/2$, resulting in

$$\bar{y} = \frac{\mathrm{b_c}}{2} - x_g \tag{5.1}$$

where x_g is the horizontal coordinate of the centroid of the section measured from the unconnected leg. The cross-section was considered as a composition of 3 elements, 2 rectangles and a quarter of a hollow circle, thus x_g is given by

$$x_g = \frac{\sum(A_i \bar{x_i})}{\sum A_i} \tag{5.2}$$

where A_i is the area of each element and $\bar{x_i}$ the horizontal coordinate of each element's centroid. It was assumed that the inner radius r_i of the hollow circle is equal to the thickness of the angle.

5.1.2. Model set I

Model set I is composed of 10 models. Initially, 6 of them were created, as shown in Table 5.3, where model M_{Exp} is the standard model, trained with all inputs and experimental data; models $M_{Exp/5}$ and $M_{Exp/3}$ are models that, based on Permutation Feature Importance (PFI), try to perform with sufficient accuracy while using only, respectively, the 5 and 3 most important inputs from model M_{Exp}; and models $M_{Exp-Num}$, $M_{Exp-Num/5}$ and $M_{Exp-Num/3}$ represent incorporation of numerical data to the training set, expanding the number of instances and aiming to evaluate if there is improvement in performance.

Table 5.3. Characteristics of the developed models

Model	No. of inputs	Training data (#Instances)
M_{Exp}	9	Exp. (123)
$M_{Exp/5}$	5	Exp. (123)
$M_{Exp/3}$	3	Exp. (123)
$M_{Exp-Num}$	9	Exp. (123) + Num. (160)
$M_{Exp-Num/5}$	5	Exp. (123) + Num. (160)
$M_{Exp-Num/3}$	3	Exp. (123) + Num. (160)

Then, the other 4 models were created, using SOFAH and SOFAH2 algorithms. Two of them using only experimental data on training set and the other two with the addition of numerical data to the training set.

There are two types of models, one trained with only experimental data and one trained with experimental and numerical data. When the training set comprehends only experimental data, the 154 instances were divided in a 4:1 proportion, with 80% (123 instances) composing the training set and 20% (31 instances) the test set. For the models trained with both experimental and numerical data, the same 31 instances were used as test set, resulting in 283 instances for training. K-fold cross-validation was done with $k = 5$ folds in all the developed models. All the partitions were done randomly and, in the same way of all random processes of the current work, following the random seed 42.

5.1.3. Model set II

Model set II is composed of models that incorporate predictions from standards' equations. SOFAH algorithm was not used in the models of this set. With the dataset information presented in section 5.1.1, models $M_{Exp-AISI}$ and $M_{Exp-AISI-EC}$ were trained on collected experimental data added to predictions given by equations from AISI standards and Eurocode 3, as shown in Table 5.4.

Table 5.4. Characteristics of the developed models that incorporate standards predictions

Model	No. of inputs	Training data (#Instances)
$M_{Exp-AISI}$	9	Exp. (123) + AISI (123)
$M_{Exp-AISI-EC}$	9	Exp. (123) + AISI (123) + Eurocode (123)

Additionally, based on the main dataset, a new one was created to compare performance outside of the training data range on developed ML model, AISI and Eurocode standards equations. Each instance of the new alternative dataset consists of mostly randomly sampled data. Connected and unconnected leg's width, respectively, b_c and b_d, were sampled following uniform distribution of range 103 to 200 mm, extrapolating the range of the experimental data, that goes up to 102 mm. The number of bolt lines n_{lines} was chose as 2 for all the instances, because the experimental dataset lacked in instances with $n_{lines} = 2$, representing only 18.2% of the set. The number of holes per line n_{holes} was randomly chosen to assume the values of 1 or 4, because these values also were lacking in the experimental dataset, representing together only 17.5%. Variables t, L and F_u were randomly sampled from their respective unique values observed in the experimental dataset. Finally, $\bar{x}$ and $\bar{y}$ were recalculated based on the values of b_c, b_d and t.

5.2. RESULTS AND DISCUSSION

5.2.1. Data exploratory analysis

For better understanding of the constructed dataset, variables statistics of experimental and numerical data are shown respectively in Table 5.5 and Table 5.6, showing minimum, maximum, median, mean, Standard Deviation (SD), skewness, Coefficient of Variation (CV), lower quartile (Q1) and upper quartile (Q3).

Table 5.5. Variables statistics of collected experimental data

Variable	Units	Min	Max	Median	Mean	SD	Skew	CV	Q1	Q3
b_c	mm	38.10	102.00	60.00	67.99	21.41	0.22	0.31	50.00	80.00
b_d	mm	38.10	102.00	60.00	67.18	22.62	0.34	0.34	50.00	82.50
t	mm	1.07	3.91	3.00	2.64	0.78	-0.48	0.30	2.25	3.00
n_{holes}	-	1.00	4.00	2.00	2.50	0.78	0.80	0.31	2.00	3.00
n_{lines}	-	1.00	2.00	1.00	1.18	0.39	1.67	0.33	1.00	1.00
$\bar{x}$	mm	6.45	34.90	15.60	18.03	7.91	0.35	0.44	10.60	26.10
L	mm	12.70	191.00	76.20	72.99	32.64	0.91	0.45	40.00	80.00
$\bar{y}$	mm	8.62	24.88	13.70	15.21	4.77	0.64	0.31	11.54	18.38
F_u	MPa	316.00	630.00	502.00	504.20	88.92	-0.68	0.18	463.00	580.00
T_{exp}	kN	15.97	179.14	85.33	85.42	36.25	0.31	0.42	61.68	107.16

Table 5.6. Variables statistics of collected numerical data

Variable	Units	Min	Max	Median	Mean	SD	Skew	CV	Q1	Q3
b_c	mm	70.00	150.00	100.00	103.81	17.48	1.20	0.17	100.00	100.00
b_d	mm	50.00	130.00	100.00	96.19	17.48	-1.20	0.18	100.00	100.00
$\bar{x}$	mm	7.40	43.70	26.60	25.38	7.70	-0.40	0.30	26.60	26.60
L	mm	33.87	76.20	38.10	46.01	13.56	1.24	0.29	38.10	50.80
$\bar{y}$	mm	13.70	38.70	23.70	24.14	5.21	0.97	0.22	23.70	23.70
T_{obs}	kN	77.96	137.25	110.36	108.02	12.42	-0.38	0.11	99.52	116.25

Note: In all instances, thickness, number of lines, number of holes per line and ultimate tensile strength of the steel material are equal, $t = 2.66$ mm, $n_{lines} = 2$, $n_{holes} = 2$ and $F_u = 502$ MPa respectively.

Note that the finite elements data (Fleitas *et al.*, 2020) expands the dataset to other ranges, for example, both b_c and b_d goes from 38.10 mm to 102.00 mm in the experimental set, as in the numerical one, simulations were performed with b_c up to 150 mm and b_d up to 130 mm.

No single-bolted angle was used in the dataset, only 4 angles presented 1 hole per bolt line, as they all have 2 bolt lines. In these 4 angles, the length of connection *L* was considered to be the same as the nominal bolt diameter, 12.7 mm,

For visualization of the experimental dataset and of the correlation between its variables, a scatterplot matrix is presented in Figure 5.1, with histograms on main diagonal. Figure 5.2 shows Pearson Correlation Matrix of all variables used. Based on the matrix, it can be affirmed that input variables $\bar{x}$, b_d, b_c and $\bar{y}$ show high correlation between them, as expected, because of geometry proportion and because b_c and b_d are both used for evaluation of $\bar{x}$ and $\bar{y}$ values. The values are $\bar{x} - b_d$ (0.94), $\bar{y} - b_d$ (0.87) and $\bar{y} - b_c$ (0.84), as the correlations are all

positive, a relation that is again expected, because of geometry proportion. Furthermore, input variables t, b_c and $\bar{y}$ show high correlation with the output variable T_{obs}, indicating that these variables may present information that helps predict T_{obs}. The values are $t - T_{obs}$ (0.76), $b_c - T_{obs}$ (0.62) and $\bar{y} - T_{obs}$ (0.60).

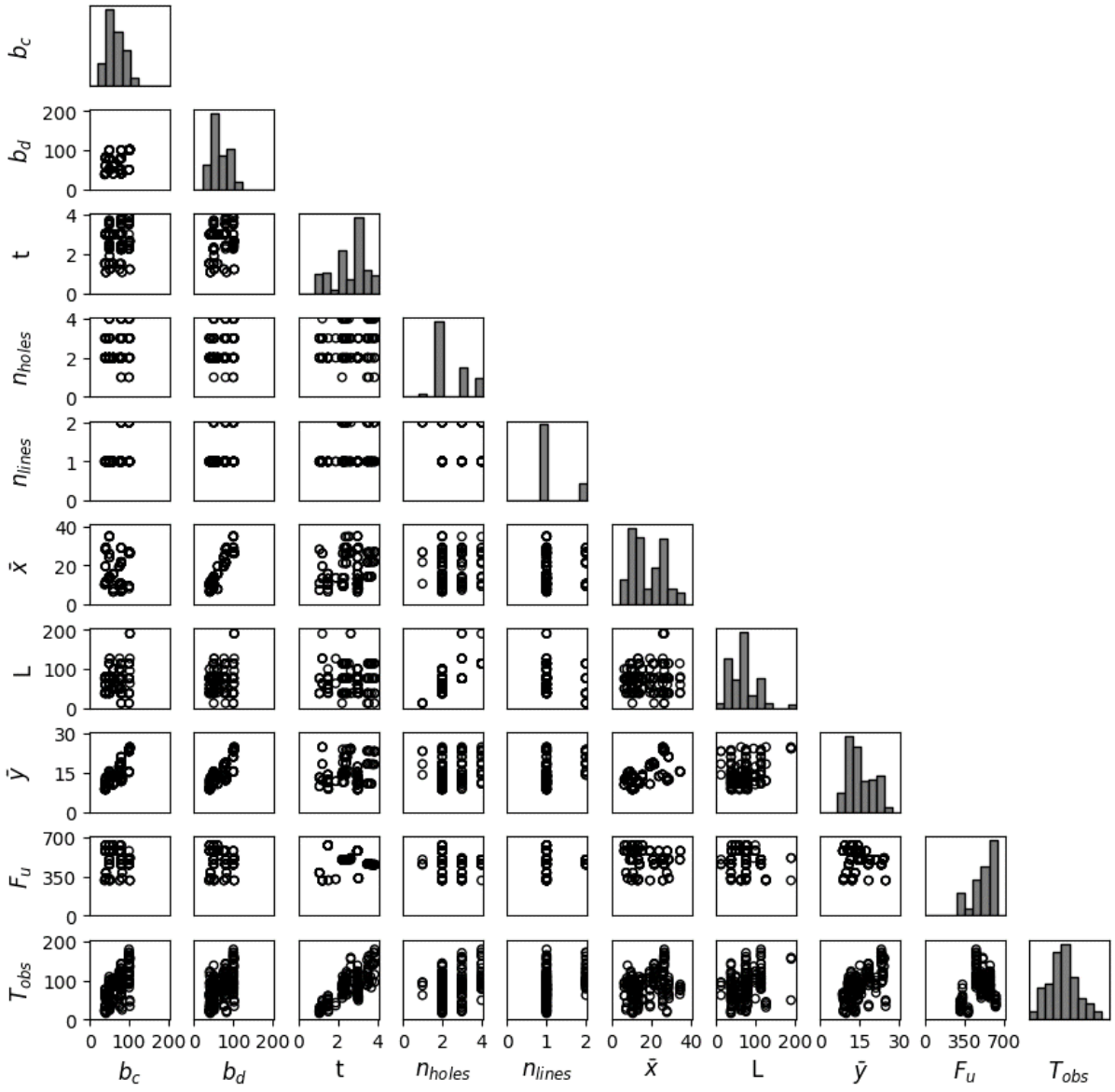


Figure 5.1. Scatterplot matrix of the experimental dataset

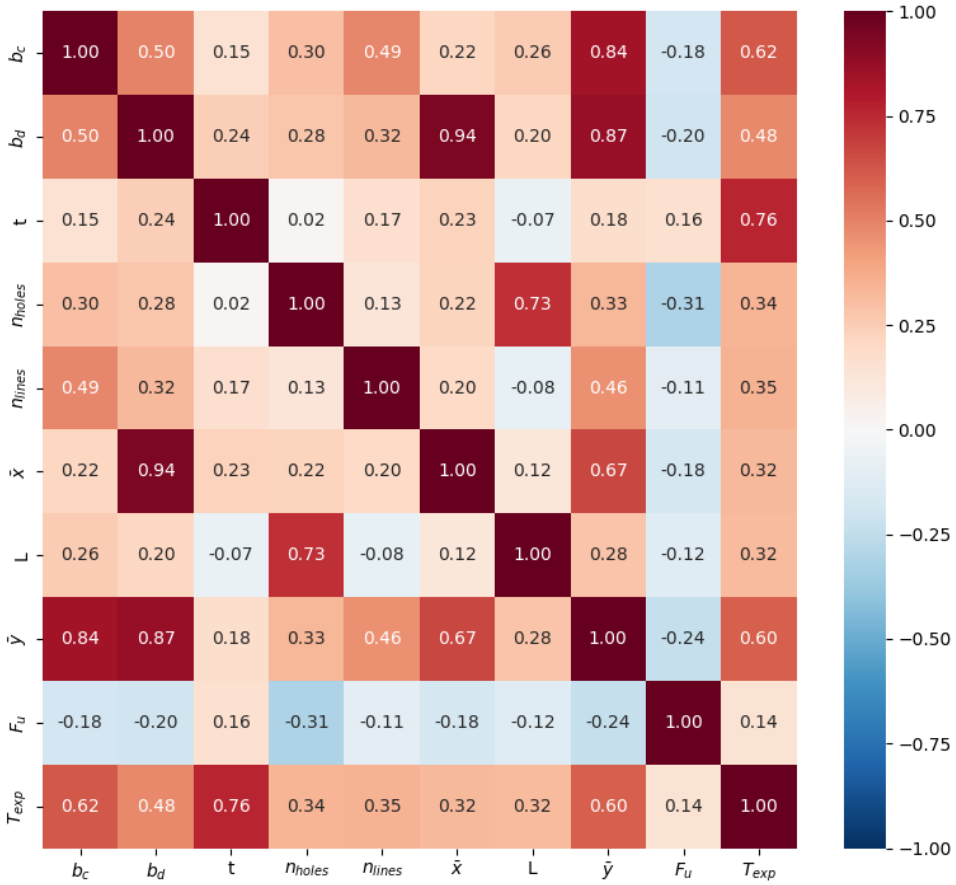


Figure 5.2. Pearson Correlation Matrix of the experimental dataset

5.2.2. Permutation feature importance and feature selection for model set I

After fitting model M_{Exp}, PFI was implemented with $N = 500$, with the increases in MSE after permutation (ΔMSE) shown as a box plot in Figure 5.3 and mean and Standard Deviation (SD) of these values shown in

Table 5.7, ordered from most to least important variable. Higher value of mean indicates that the analyzed input feature significantly contributes to lower MSE on the model. In reason of the mentioned results: models $M_{Exp/5}$ and $M_{Exp-Num/5}$ are composed of t, b_c, F_u, L and n_{holes} and models $M_{Exp/3}$ and $M_{Exp-Num/3}$ use t, b_c and F_u.

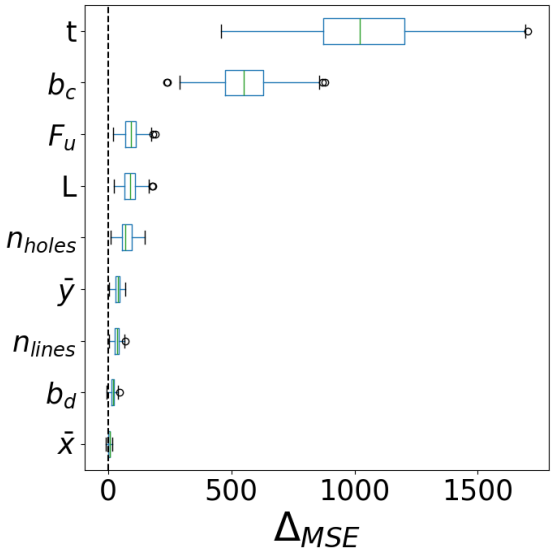


Figure 5.3. PFI box plot for increase in MSE (ΔMSE)

Table 5.7. Mean and SD of ΔMSE after PFI

Features	Mean	SD
t	1035.36	230.17
b_c	552.28	117.45
F_u	93.28	31.13
L	89.26	28.75
n_{holes}	75.78	24.84
$\bar{y}$	38.12	11.86
n_{lines}	35.95	11.68
b_d	19.07	8.86
$\bar{x}$	4.49	5.68

Steel angle thickness t and width of the connected leg b_c were clearly the most important features for net section resistance prediction in the developed model. This result is aligned with the fact that the 2 variables were the ones of greater correlation with the output. On the other side, F_u was the third most important variable and showed the lowest Pearson's correlation with T_{exp}. A reasonable explanation to this is that Pearson's correlation coefficient measures only linear correlation between variables, thus F_u may be nonlinearly correlated with T_{exp}.

5.2.3. Model set I

Since the SBA dataset is very small, containing only a few hundred instances, all the models

developed here took less than 30 seconds in training. The baseline methodology and SOFAH algorithm took a maximum of 2.5 hours in HPO. SOFAH2, since it performs a second BO after SOFAH's own BO, took 4 hours and 21 minutes in HPO.

To compare models M_{Exp}, $M_{Exp/5}$, $M_{Exp/3}$, $M_{Exp-Num}$, $M_{Exp-Num/5}$ and $M_{Exp-Num/3}$ some regression evaluation metrics were calculated in training and testing stages: RMSE, R^2 and MAPE. All the mentioned metrics results are shown in Table 5.8.

Table 5.8. Regression evaluation metrics of developed models in training and test

Performance metrics	$\mathbf{M_{Exp}}$	$\mathbf{M_{Exp/5}}$	$\mathbf{M_{Exp/3}}$	$\mathbf{M_{Exp-Num}}$	$\mathbf{M_{Exp-Num/5}}$	$\mathbf{M_{Exp-Num/3}}$
RMSE - Training	3.44	4.39	10.09	4.59	6.89	11.44
RMSE - Test	6.37	7.70	13.81	6.87	5.43	13.18
R^2 - Training	0.99	0.99	0.92	0.97	0.94	0.83
R^2 - Test	0.97	0.96	0.86	0.97	0.98	0.88
MAPE - Training	3.61%	3.93%	9.43%	3.82%	5.13%	9.65%
MAPE - Test	6.11%	7.78%	10.84%	6.61%	5.70%	10.53%

As observed in Table 5.8, the standard model, M_{Exp}, using the original set of 9 features and only experimental data, was the best performing model among the models trained only on experimental data, showing R^2 of 0.97, RMSE of 6.37 and MAPE of 6.11% in test set. Respectively, these results can be interpreted as, for unseen instances, 97% of the variation of T_{obs} was accounted for by the independent variables, the SD of prediction errors was 6.37 kN, and, on average, each instance showed a relative error of 6.11%. When comparing the models, all of them showed good performance, with R^2 greater than 0.85 and MAPE lower than 11.00%, but $M_{Exp-Num/5}$, a model with only 5 of the 9 available features and with numerical and experimental data in the training set, was the one of best metrics, showing the lowest values for the error metrics, MAPE and RMSE, and the higher R^2. The addition of numerical data to the training set provided similar results when using all 9 inputs and 3 inputs, but, when using 5 inputs, it improved the results and provided the best performing model.

For visualization of predictive accuracy in all of the points of the test set, Figure 5.4 shows predicted values plotted against observed values of the target variable, presenting side-by-side

comparisons of M_{Exp} and $M_{Exp-Num}$, $M_{Exp/5}$ and $M_{Exp-Num/5}$, and $M_{Exp/3}$ and $M_{Exp-Num/3}$. In the plot, the closer the point is to the $y = x$ diagonal line, the more accurate was its prediction. The dashed lines show margins of error of $+10\%$ and -10%.

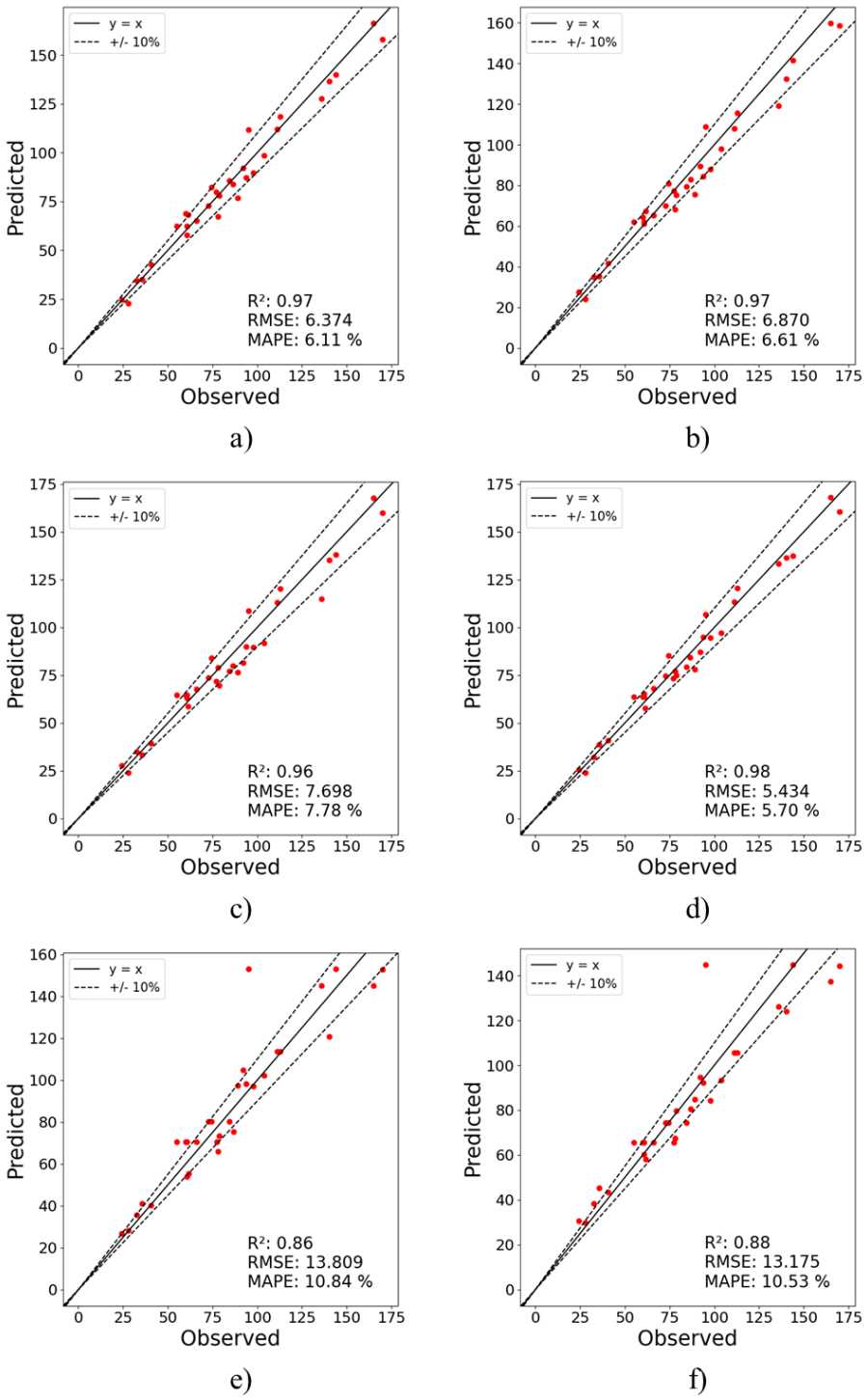


Figure 5.4. Predicted vs. Observed values for models: a) M_{Exp}; b) $M_{Exp-Num}$; c) $M_{Exp/5}$; d) $M_{Exp-Num/5}$; e) $M_{Exp/3}$; f) $M_{Exp-Num/3}$

Comparing a) and b), although very similar, on average M_{Exp} presents predictions slightly more accurate than $M_{Exp-Num}$. Comparing d) with all the others, it is visually clear that $M_{Exp-Num/5}$ really shows the best accuracy when analyzing all the test points. Investigating e) and f), both models $M_{Exp/3}$ and $M_{Exp-Num/3}$ do not show accuracy as bad as the metrics indicated, because there is a single point of extremely bad prediction that pulled MAPE and RMSE towards high values. The mentioned point corresponds to one of four highlighted angles (the other three are on training set) that had only 1 hole per line and 2 lines, where length of connection L was considered equal to the nominal bolt diameter of 12.7 mm. In short, the result is expected, since both models of 3 inputs used only t, b_c and F_u, they had insufficient information to predict this specific angle of single bolt in longitudinal direction.

Model M_{Exp} predictions were then compared to the ones given by equations from AISI (2016) and Eurocode 3 (2005a, 2005b) standards; and equations proposed by Fleitas *et al.* (2020), Paula *et al.* (2008) and Teh and Gilbert (2013). The comparison between predictions and observed values was done through some performance metrics evaluated on test set, MSE, RMSE, R^2 and MAPE and its results are shown in Table 5.9.

Table 5.9. Comparison of evaluation metrics of each prediction and observed values

Models/Equations	**MSE - Test**	**RMSE - Test**	**R^2 - Test**	**MAPE - Test**
Model M_{Exp}	40.63	6.37	0.97	6.11%
Model $M_{Exp-Num/5}$	29.53	5.43	0.98	5.70%
Model $M_{Exp-AISI}$	101.92	10.10	0.93	9.89%
Model $M_{Exp-AISI-EC}$	341.70	18.49	0.75	13.28%
AISI (Eq. (2.31))	277.65	16.66	0.80	14.85%
Eurocode 3 (Eqs. (2.32)(2.33))	2053.58	45.32	-0.48	29.81%
Paula *et al.* (Eq. (2.35))	316.30	17.78	0.77	17.60%
Teh and Gilbert (Eq. (2.36))	160.72	12.68	0.88	13.20%
Fleitas *et al.* (Eq. (2.37))	298.80	17.29	0.79	15.53%

Note: The values compared with the predictions of each model/equation are the observed values from all samples of the test set.

As observed in Table 5.9, AISI Standards clearly provide the best evaluation metrics among the standards. The equations proposed by some relevant works of the area in general perform

better than the equations from standards, specially Teh and Gilbert (2013), that shows the highest R^2, 0.88, and the lowest RMSE, 12.68. This behavior is expected, because these equations have the advantage of already having seen some of the instances from the test set, since they were used for regression of the analyzed equations. Equation (2.36), proposed by Teh and Gilbert (2013), used approximately one third of the experimental data for its regression. When compared to the ML model, model M_{Exp} shows outstanding accuracy, with R^2 of 0.97 and RMSE of 6.37 on test set. The bad performance of Eurocode equation may be explained by the specific and small test set analyzed, as standards equations are effective in wider domain. Figure 5.5 shows predicted values plotted against observed values of the target variable, comparing results of AISI (2016), Eurocode (2005a, 2005b), Paula *et al*. (2008), Teh and Gilbert (2013) and Fleitas *et al*. (2020).

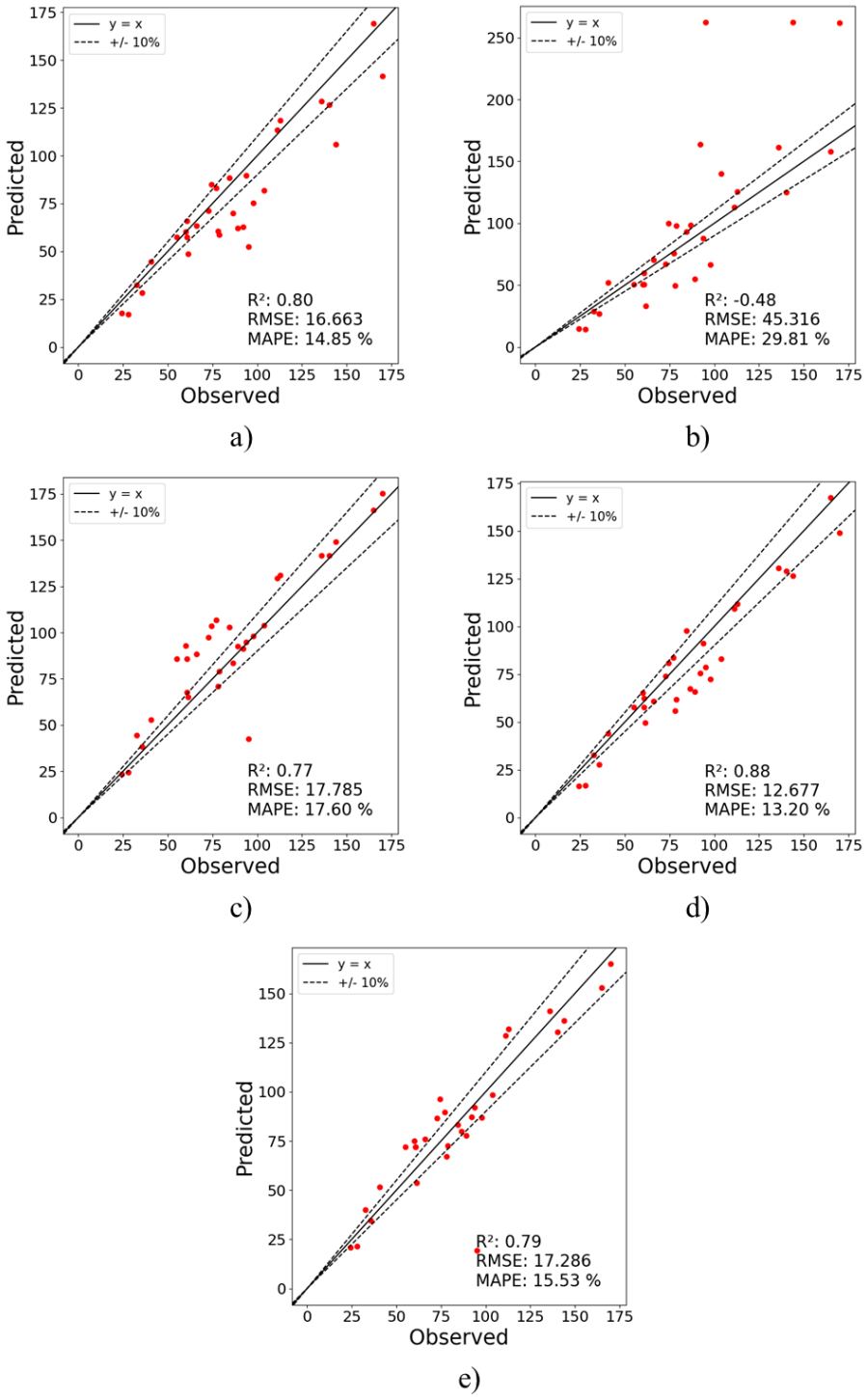


Figure 5.5. Predicted vs. Observed values for: a) AISI; b) Eurocode 3; c) Paula *et al.*; d) Teh and Gilbert; e) Fleitas *et al.*

Observing Figure 5.5 a) and d), it is visually clear that AISI and Teh and Gilbert equations are generally conservative, presenting most of the results below the diagonal lines, that is, predicted values lower than observed ones. As seen in b), Eurocode provided inaccurate predictions for

considerable portion of the 31 instances of the test set. In c) and e), it is plausible to say that Paula *et al.* and Fleitas *et al.* proposed good equations, but they showed one point of very bad prediction and considerable amount of predictions above the diagonal lines, thus being not favorable to safety. This point is one of only four instances that presented a single bolt per line, where it was considered that the length of connection L is equal to the bolt diameter (12.7 mm). Comparing these results to the best ML model in Figure 5.4 d), it clear how much more accurate $M_{Exp-Num/5}$ is.

Figure 5.6 shows a comparison between predictions on test set of the best performing standard, AISI, proposed equation, Teh and Gilbert, and model, $M_{Exp-Num/5}$, highlighting if they are favorable to safety or not, as a margin of error of +/- 10% is considered.

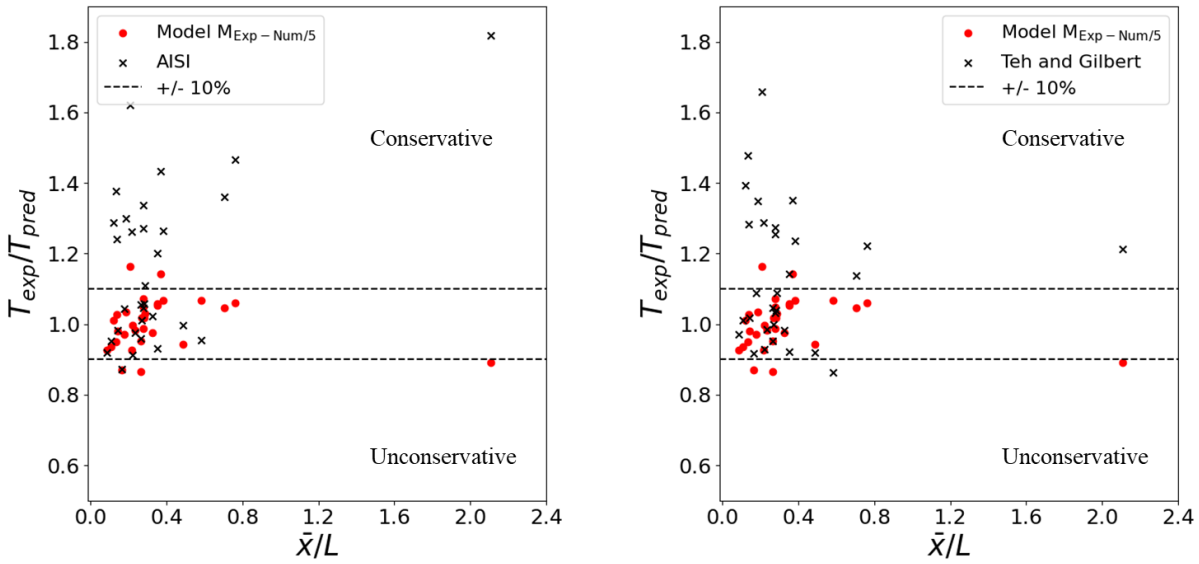


Figure 5.6. Comparison of observed/predicted on test set vs. $\bar{x}/L$: a) AISI; b) Teh and Gilbert

As shown, it can be said that: AISI predictions are really mostly conservative; Teh and Gilbert predictions, although more accurate, are also mostly conservative; Model $M_{Exp-Num/5}$ is much more accurate than the analyzed equations, presenting almost all predictions between the 10% error margin. The outlier of $\bar{x}/L$ greater than 2.0 is one of only four instances that presented a single bolt per line, where it was considered that the length of connection L is equal to the bolt diameter (12.7 mm). This sample showed the least accurate and most conservative prediction of the AISI equation.

5.2.4. Model set I with SOFAH algorithm

The best performing model of only experimental training data, M_{Exp} and best one with addition of numerical data to the training set, $M_{Exp-Num/5}$ were compared with their SOFAH and SOFAH2 counterparts, that is, M_{Exp}^{SOFAH}, M_{Exp}^{SOFAH2}, $M_{Exp-Num}^{SOFAH}$ and $M_{Exp-Num}^{SOFAH2}$. The results of evaluated performance metrics are shown in Table 5.10.

Table 5.10. Performance metrics of SOFAH models

Performance metrics	$\mathbf{M_{Exp}}$	$\mathbf{M_{Exp}^{SOFAH}}$	$\mathbf{M_{Exp}^{SOFAH2}}$	$\mathbf{M_{Exp-Num/5}}$	$\mathbf{M_{Exp-Num}^{SOFAH}}$	$\mathbf{M_{Exp-Num}^{SOFAH2}}$
RMSE - Training	3.44	2.42	3.19	6.89	3.96	3.98
RMSE - Test	6.37	6.15	6.70	5.43	5.27	4.77
R^2 - Training	0.99	1.00	0.99	0.94	0.98	0.98
R^2 - Test	0.97	0.97	0.97	0.98	0.98	0.98
MAPE - Training	3.61%	2.60%	3.12%	5.13%	3.06%	3.30%
MAPE - Test	6.11%	5.88%	6.13%	5.70%	5.44%	5.29%

As seen, all SOFAH-based models resulted in better metrics in both training and test set, indicating better fit to the data and also higher generalization capacity. In experimental-only models, M_{Exp}^{SOFAH}, compared to M_{Exp}, achieved the best results, with a lower RMSE and MAPE. When adding numerical data to the training set, the previous best performing model, $M_{Exp-Num/5}$, was improved by $M_{Exp-Num}^{SOFAH}$, and even further by $M_{Exp-Num}^{SOFAH2}$, resulting in the best metrics yet, RMSE of 4.77 and MAPE of 5.29% on test set. Predicted vs. observed values plot of the best performing model, $M_{Exp-Num}^{SOFAH2}$, is shown in Figure 5.7. It is possible to see how all the points provided very accurate predictions.

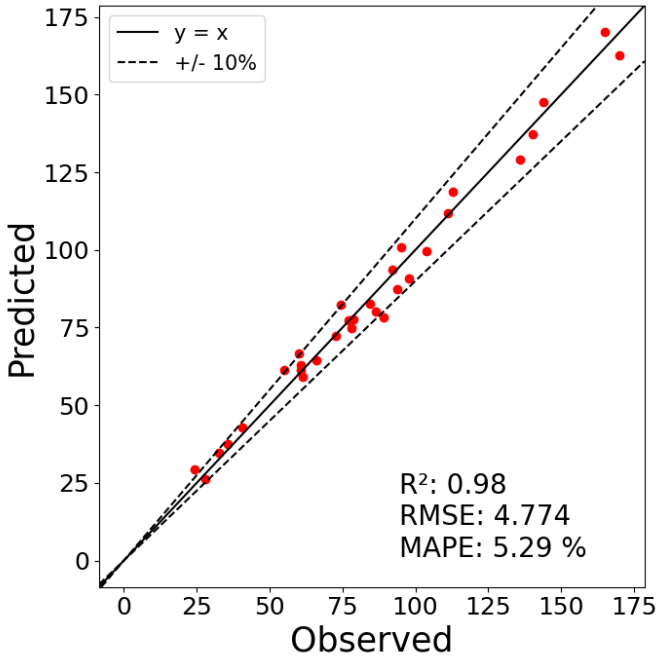


Figure 5.7. Predicted vs. observed values of model $M_{Exp-Num}^{SOFAH2}$

Also, in Figure 5.8 and Figure 5.9, it is shown the optimal feature subsets found by SOFAH algorithm in M_{Exp}^{SOFAH} and $M_{Exp-Num}^{SOFAH}$ respectively. The former is composed of 16 variables out of the 81 obtained by the feature augmentation performed by the algorithm. The latter is composed of only 11 variables, where 5 of them include the length of connected leg b_c, the second most important variable of the baseline model, as shown previously in Figure 5.3.

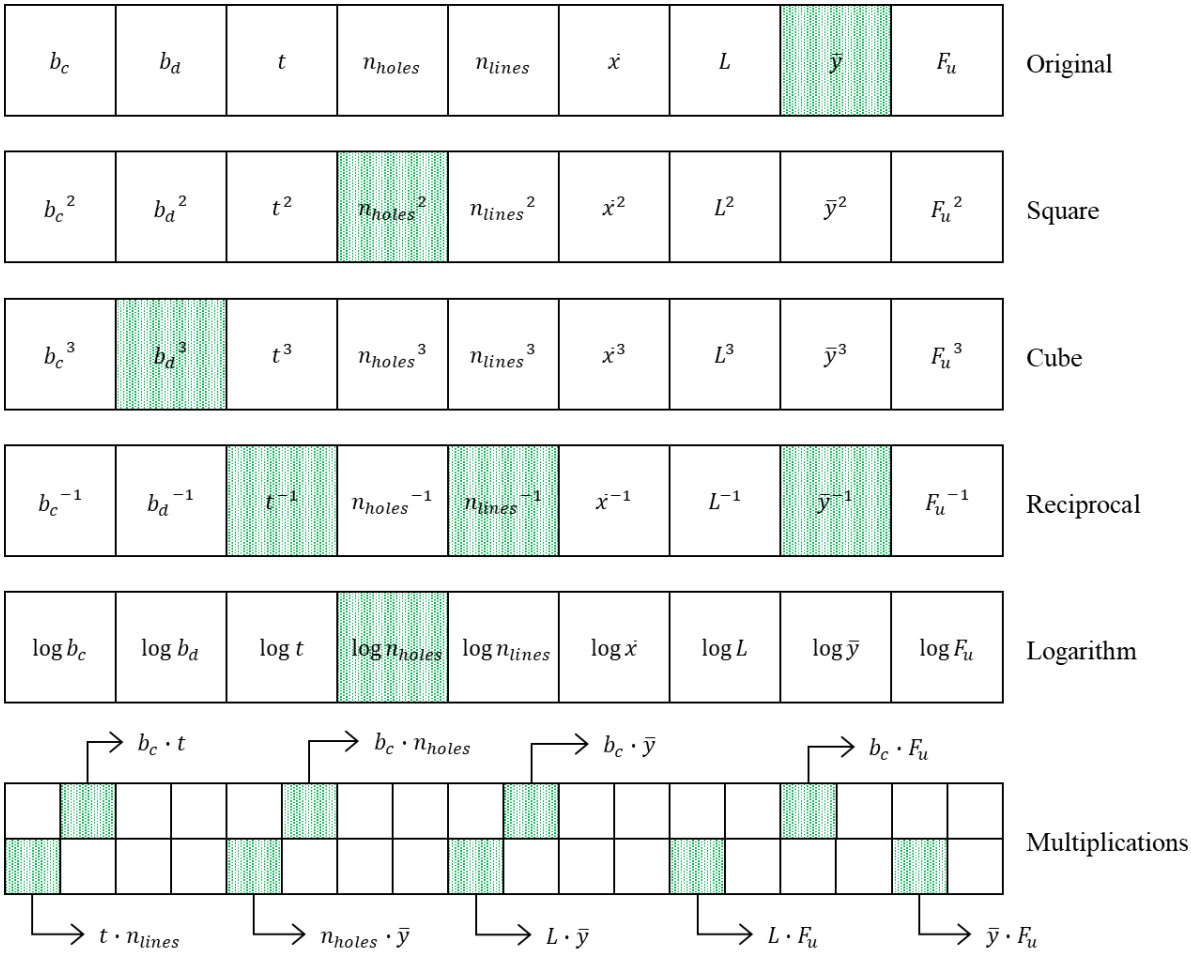


Figure 5.8. Optimal feature subset of M_{Exp}^{SOFAH}

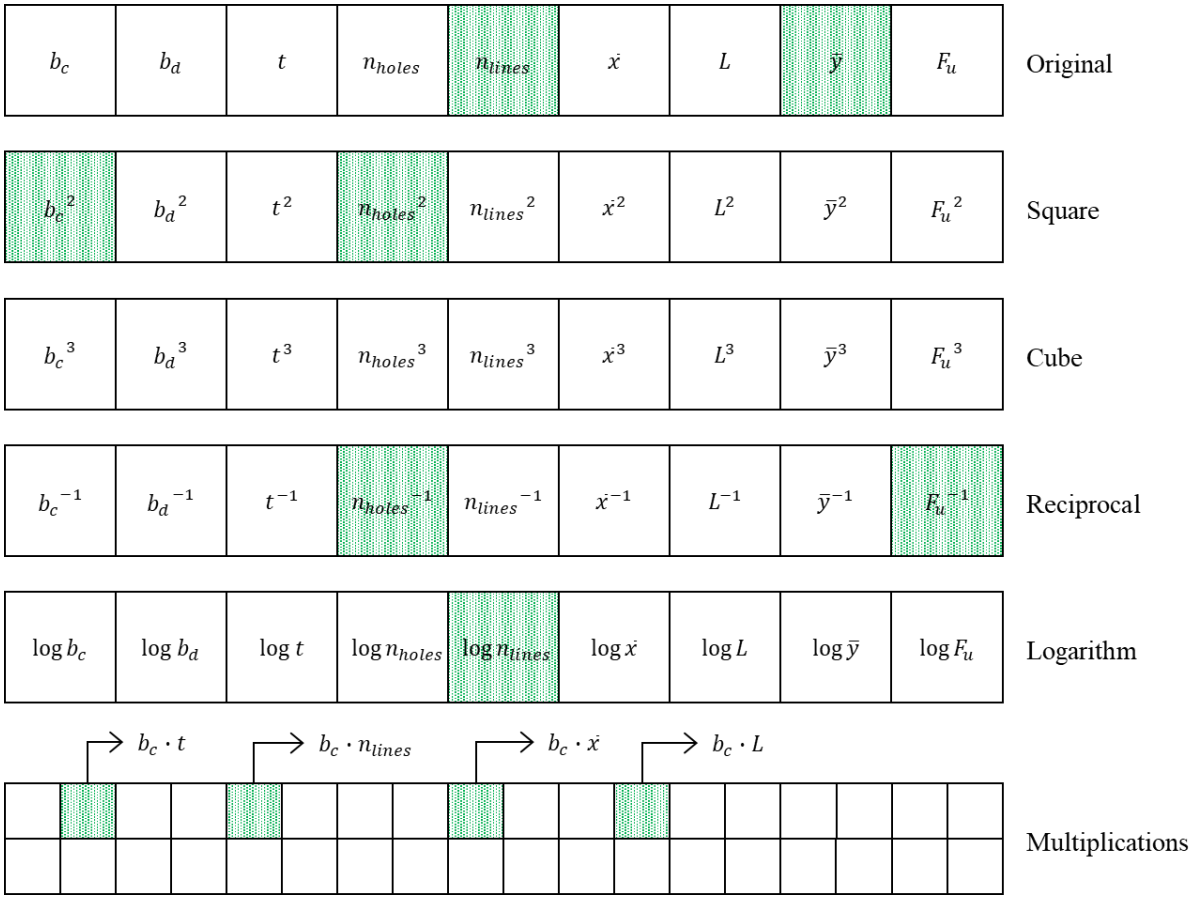


Figure 5.9. Optimal feature subset of $M_{Exp-Num}^{SOFAH}$

5.2.5. Model set II

Aiming to develop ML models favorable to safety and capable of performing predictions in a wider domain, models $M_{Exp-AISI}$ and $M_{Exp-AISI-EC}$ were built adding to the training set predictions from the equations given by North American and European standards. Their performance metrics, MSE, RMSE, R^2 and MAPE, were evaluated on the 31 instances of observed experimental data that compose the test set and are shown in Table 5.11.

Table 5.11. Evaluation metrics of model set II

Models/Equations	**MSE - Test**	**RMSE - Test**	**R^2 - Test**	**MAPE - Test**
Model $M_{Exp-AISI}$	101.92	10.10	0.93	9.89%
Model $M_{Exp-AISI-EC}$	341.70	18.49	0.75	13.28%

Comparing the results with metrics of Table 5.9, it can be said that both models $M_{Exp-AISI}$ and $M_{Exp-AISI-EC}$ presented worse values for all metrics than the standard ML models M_{Exp},

$M_{Exp/5}$ and $M_{Exp-Num/5}$. However, model $M_{Exp-AISI}$ showed better metrics than all other equations and models, thus representing an interesting more conservative ML model that incorporated standards' predictions.

The two models from model set II were used to perform predictions on an alternative dataset, outside of the range of training data. Figure 5.10 shows plots of T_{AISI}/T_{pred} vs. $\bar{x}/L$ for predictions of models $M_{Exp-Num/5}$, M_{Exp}, $M_{Exp-AISI}$, and $M_{Exp-AISI-EC}$ on the alternative dataset, comparing results two by two.

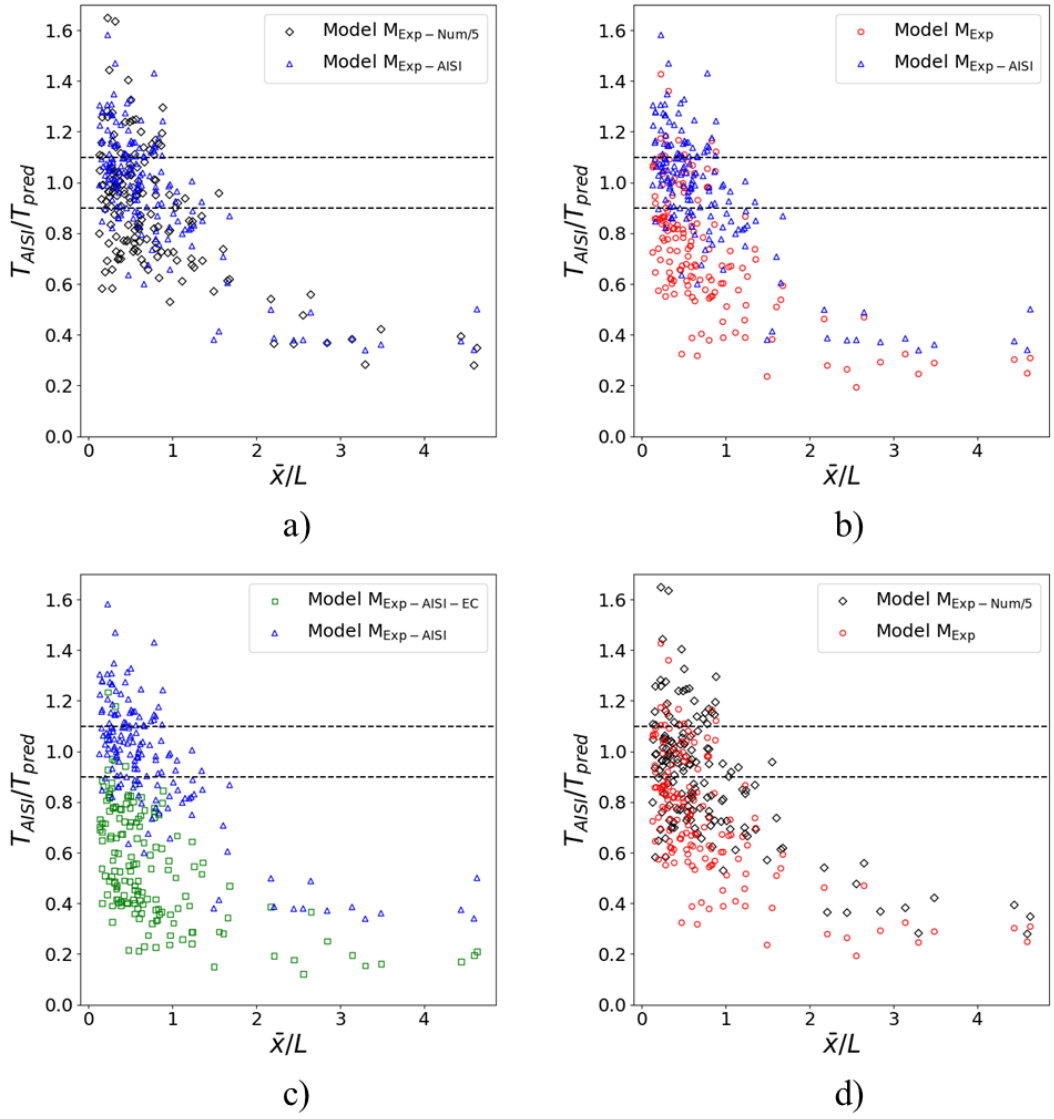


Figure 5.10. Comparison of T_{AISI}/T_{pred} between: a) $M_{Exp-Num/5}$ and $M_{Exp-AISI}$; b) M_{Exp} and $M_{Exp-AISI}$; c) $M_{Exp-AISI-EC}$ and $M_{Exp-AISI}$; d) $M_{Exp-Num/5}$ and M_{Exp}

As discussed on the analysis of Figure 5.6 a), AISI equation is generally conservative and even more in high values of $\bar{x}/L$, predicting low values of resistance. Based on that, the behavior shown in Figure 5.10 of decrease in T_{AISI}/T_{pred} for higher values of $\bar{x}/L$ is expected. Based on Figure 5.10 c) Model $M_{Exp/AISI/EC}$ provided the most conservative predictions, with almost all the points below the horizontal lines, that is, predicting smaller values than AISI in almost all simulated instances. Comparatively, in a), models $M_{Exp/AISI}$ and $M_{Exp-Num/5}$ presented the most instances between and above the 10% margins of error. In b) and d), model M_{Exp} clearly predicted lower values than these two models, with points more distant from the horizontal lines, that is, higher error when compared to AISI predictions.

5.2.6. Comparison of model sets I and II with AISI equation

Another comparison of the same metrics was done, but now, instead of comparing predictions and observed values, the predictions from each equation and model were compared with AISI predictions. The evaluations were performed on the same 31 samples that compose the test set and the results are shown in Table 5.12.

Table 5.12 Comparison of evaluation metrics of each prediction and AISI predictions

Models/Equations	**MSE - Test**	**RMSE - Test**	**R^2 - Test**	**MAPE - Test**
Model M_{Exp}	257.09	16.03	0.79	17.75%
Model $M_{Exp/5}$	223.98	14.97	0.82	17.66%
Model $M_{Exp-Num/5}$	235.27	15.34	0.81	18.15%
Model $M_{Exp-AISI}$	88.38	9.40	0.93	8.64%
Model $M_{Exp-AISI-EC}$	874.89	29.58	0.29	21.31%
Eurocode 3 (Eqs. (2.32)(2.33))	3270.54	57.19	-1.64	40.38%
Paula *et al.* (Eq. (2.35))	425.11	20.62	0.66	27.68%
Teh and Gilbert (Eq. (2.36))	52.66	7.26	0.96	6.22%
Fleitas *et al.* (Eq. (2.37))	209.21	14.46	0.83	18.56%

Note: The values compared with the predictions of each model/equation are the AISI predictions from all samples of the test set.

As observed, Teh and Gilbert predictions were the closest to AISI predictions, that occurred because of the extreme similarity between both equations, (2.31) and (2.36). Model $M_{Exp/AISI}$

showed the second closest predictions, thus assuring that this model incorporated successfully AISI predictions. Note that, compared to AISI predictions, the best performing model was $M_{Exp/5}$, differing from the best performing compared to observed values, that was $M_{Exp-Num/5}$.

6. TEMPERATURE VARIATIONS ON RIGID PAVEMENT CONCRETE SLABS

According to data from the National Transport Confederation (CNT, Confederação Nacional do Transporte, in Portuguese) released in the 26th edition of the CNT Highway Survey on November 29, 2023 (CNT, 2023), road transport is responsible for moving 65% of the country's freight and 95% of its passengers. However, 67.5% of the road network is classified as average, poor or bad, while only 32.5% is classified as good or excellent. The classification considers three main characteristics of the road network: pavement, signalization and road geometry. Specifically, regarding pavement quality, 56.8% of the evaluations were average, poor or bad.

The primary objective of a pavement is to deliver a durable and functional surface tailored to specific transportation requirements. Its fundamental role is to sustain applied loads under varying seasonal and environmental conditions without experiencing deformation or cracking, as such distresses would compromise the pavement's performance and longevity (MALLICK; EL-KORCHI, 2023).

According to the Brazilian Association of Portland Cement (ABPC, Associação Brasileira de Cimento Portland, in Portuguese), the National Department of Transport Infrastructure's (DNIT, Departamento Nacional de Infraestrutura de Transportes, in Portuguese) Director of Planning and Research, engineer Luiz Guilherme Melo, noted the growing trend of using concrete pavement on federal highways managed by the agency, which currently accounts for 4.5% of the network, with the prospect of reaching 10% (ABPC, 2024). In 2024, eight concrete pavement projects were being executed, totaling 673 kilometers, with a focus on the North and Northeast regions of Brazil. These projects used modern technologies, bringing improvements to the infrastructure.

Rigid pavements, commonly constructed using concrete experience minimal deflection under load due to their inherent stiffness, in opposition to flexible (or asphalt) pavement, that deflects under traffic load. The surface layer, which directly interacts with traffic, consists of a Portland cement concrete slab. The high stiffness of the concrete slab, characterized by its elastic modulus, enables it to distribute loads over a wider area, thereby reducing the stress transferred to the underlying base and subgrade layers. While rigid pavements can be constructed directly on the subgrade, high-performance designs often incorporate a base or subbase layer to enhance structural integrity. In addition to serving as a wearing surface, the concrete slab contributes to

critical functions such as providing skid resistance, facilitating drainage, ensuring ride smoothness, mitigating noise, and offering waterproofing protection to the underlying layers (MALLICK; EL-KORCHI, 2023).

Concrete pavements are commonly known for being more durable, thus needing reduced maintenance and consequently reducing its general costs. For this reason, concrete has been used in areas with intense bus and truck traffic, such as exclusive lanes and highways. The temperature distribution along the depth of the concrete pavement, that is, the temperature profile, is extremely important for determining deformations and stresses resulting from thermal variations between the upper and lower layers of the pavement (WANG; ROESLER, 2014). In case of not considering these variations, there can be a significant impact on the behavior and structural performance of the pavement, even reaching the ultimate limit state and collapsing. In regions of high daily thermal amplitude and high solar radiation, like many countries in Brazil, the study of rigid pavement behavior relative to temperature variations becomes even more relevant.

6.1. METHODOLOGY

6.1.1. Dataset

The general methodology is shown in Figure 6.1 and consists of using the equations of heat flow theory and climate data to predict surface temperature, T_{surf}, over the span of a few years. The values obtained are used in a Finite Element Method (FEM) model in ABAQUS software to simulate the climate conditions and evaluate the temperate at varying depths of the concrete slab layer. With these results, it is possible to calculate the difference in temperature from surface to bottom of the slab, ΔT. Finally, ANN models are created with all this generated data, aiming to be used as efficient surrogate models of reduced costs. For better understanding, the described process is also shown with other details in the workflow diagram in Figure 6.2, even specifying the name of the models created.

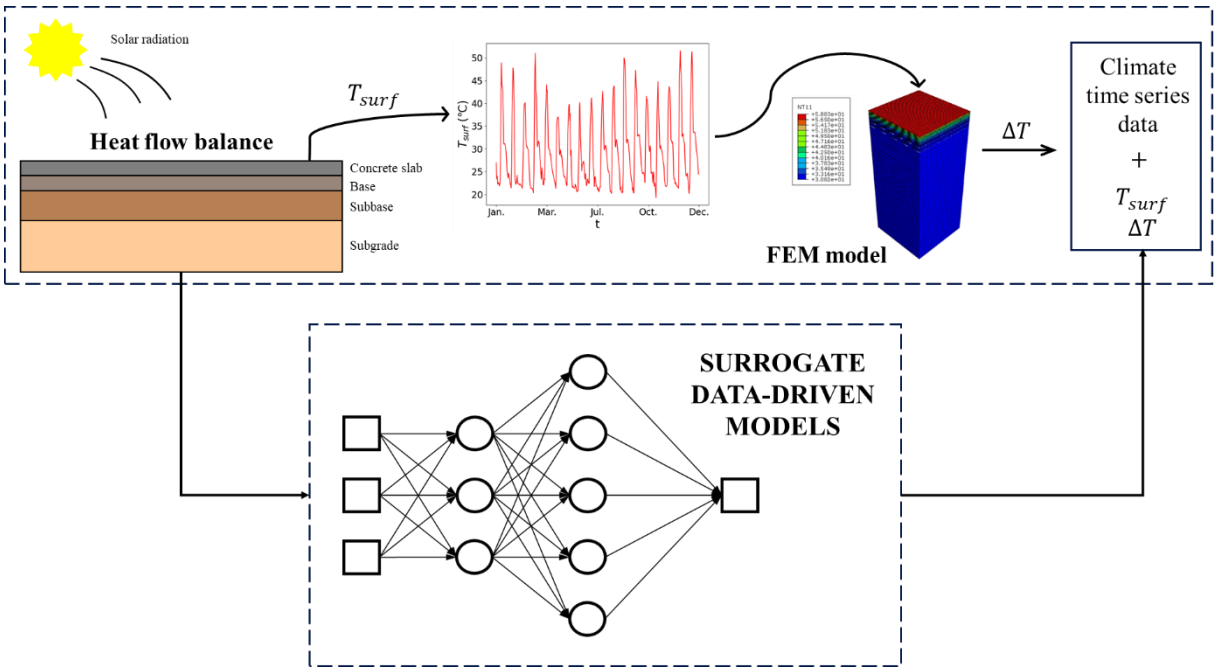


Figure 6.1. Surrogate modeling of temperature variations in concrete pavement

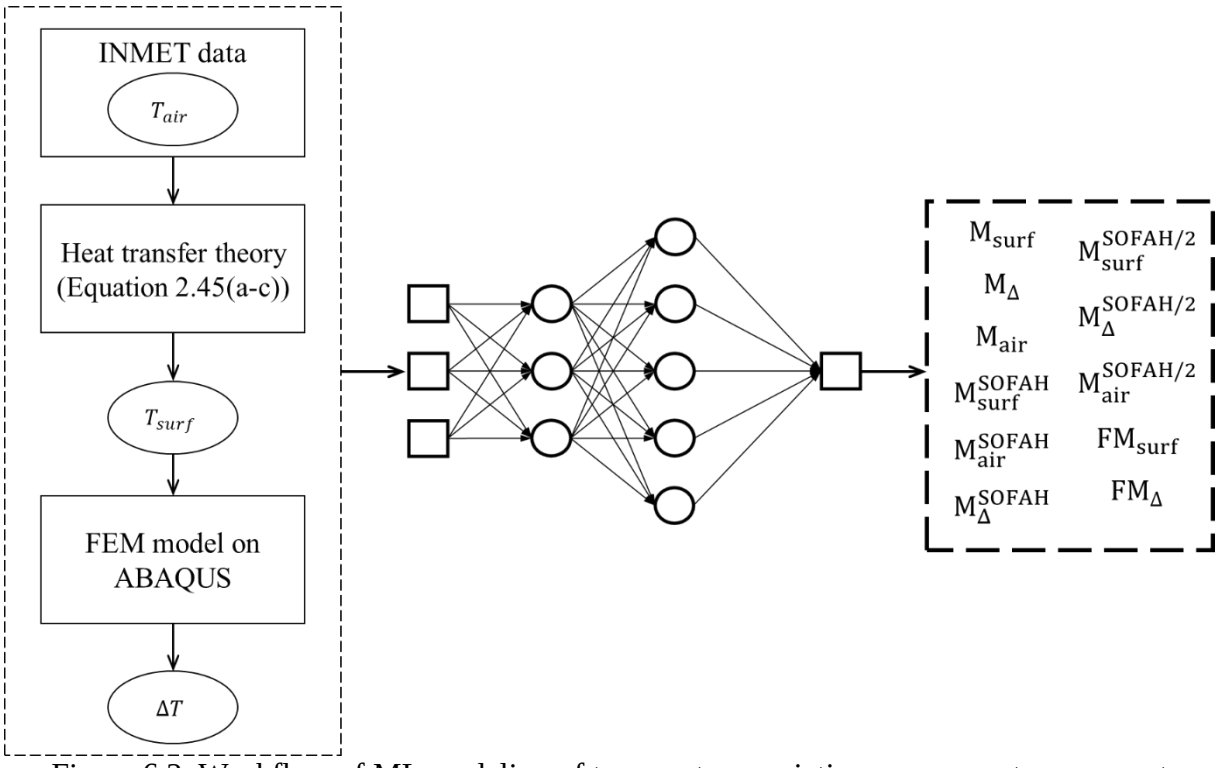


Figure 6.2. Workflow of ML modeling of temperature variations on concrete pavement

The National Institute of Meteorology (INMET, Instituto Nacional de Meteorologia, in Portuguese), a federal agency under the direct administration of the Ministry of Agriculture,

Livestock and Food Supply, publicly provides a meteorological database acquired hourly from all automatic weather stations across the country, organized by year. The dataset used in the present research corresponds to all data from station 001, located in Brasília, Brazil, situated at latitude and longitude (-15,789343, -47,925756), from year 2019 to year 2023. Since this database was measured hourly, it can be said that each non-leap year presents 8760 instances. The climate data was collected from INMET database already containing air temperature values.

Using the disposed climatic variables, the temperature T_{surf} at the surface of the concrete slab was calculated following the formulation of heat transfer theory explained in section 2.2.3 and resumed in Equation 2.45(a-c). The values of deterministic variables present in this equation are shown in Table 6.1. Some considerations about the variables are: α_a for concrete material goes from 0.65 to 0.80, thus being used its average of 0.725; emissivity value ε was considered the same as solar absorptivity α_a; k_c was considered as 1.42 W/m°C for concrete; R_0 and σ_0 are constants with respective units W/m² and W/m²°K^4; and τ_a ranges from 0.81 on clear days to 0.62 on cloudy ones.

Table 6.1. Deterministic variables used for pavement surface temperature prediction

Variable	Value
α_a	0.725
ε	0.725
k_c	1.42
R_0	1394
σ_0	$5.68 \cdot 10^{-8}$
τ_a	0.81
ε_a	0.70

The values of T_{surf} were used to determine the temperature at different depths of the slab through simulations conducted in the ABAQUS software by Teche and Evangelista Jr. (2025). The authors employed 5 pavement configurations, varying the thickness h_1 of the concrete slab and thickness h_2 of the base, as shown in Table 6.2. The objective of these configurations was to include h_1 and h_2 in the models, achieving regression models that are not limited to a single specification of pavement system. Additionally, the material properties used for each pavement layer in the simulations are detailed in Table 6.3.

.

Table 6.2. Thickness of each layer and respective configuration

Configuration	Concrete slab (cm)	Base (cm)	Subbase (cm)
I	25	10	300
II	23	10	300
III	18	20	300
IV	25	20	300
V	20	20	300

Table 6.3. Material properties used in ABAQUS simulations

Properties	Concrete pavement	Roller-compacted concrete base	Soil
Density (kg/m³)	2291	2400	1700
Thermal conductivity (W/mºC)	1.42	2.10	1.10
Specific heat (J/mºC)	957.00	920.48	750.00

Using simulation results, ΔT was calculated, the difference between surface temperature T_{surf} and bottom temperature of the concrete slab. The dataset assembled all variables disposed in Table 6.4.

Table 6.4. Identification and description of each variable of the dataset

Feature	Unit	Description
D	-	Current day
H	-	Current hour
T_{air}	ºC	Air temperature
P_{atm}	mbar	Atmospheric pressure
RH	%	Relative humidity
ff	m/s	Wind speed
T_{surf}	ºC	Concrete slab surface temperature
ΔT	ºC	Temperature difference between surface and bottom of the concrete slab

As observed in Table 6.2, the only difference between the configurations are the thickness of the layers, so, all climatic variables show the same values from one configuration to another, D, H, T_{air}, P_{atm}, RH and ff. Furthermore, T_{surf} is not affected by the different configurations, as it is calculated with latitude coordinate and air temperature, two variables that are also not

affected. The only variable affected by different thickness of pavement layers is ΔT.

6.1.2. Model set I

Initially, model set I was built, consisting of 9 distinct models that use only configuration I data (Table 6.2). The inputs and output of 3 of these models are described in Table 6.5, where T_{surf}^{pred} corresponds to predicted values of T_{surf}, that is, model M_{Δ} uses the predictions from $\mathrm{M}_{\mathrm{surf}}$, $\mathrm{M}_{\Delta}^{\mathrm{SOFAH}}$ from $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH}}$, and $\mathrm{M}_{\Delta}^{\mathrm{SOFAH2}}$ from $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH2}}$. The inputs of the rest of the models were defined only after executing SOFAH algorithm.

Table 6.5. Inputs and output of model set I (configuration I)

Model	Inputs	Output
$\mathrm{M}_{\mathrm{air}}$	D, H, P_{atm}, RH, ff	T_{air}
$\mathrm{M}_{\mathrm{surf}}$	$D, H, P_{atm}, RH, ff, T_{air}$	T_{surf}
M_{Δ}	$D, H, P_{atm}, RH, ff, T_{air}, T_{surf}^{pred}$	ΔT

It is fundamental to state that in all the 9 models that compose models set I, the training set is composed of 80% of the data from 01/04/2019 to 12/31/2020, as the other 20% were used as the test set. This means that model set I was trained on 13977 instances and tested on 3495 instances. All the data from 01/04/2021 to 12/31/2023 was exclusively used for inference on model set II.

6.1.3. Model set II

Model set II was trained, comprehending 2 new models, $\mathrm{FM}_{\mathrm{surf}}$ and FM_{Δ}, that use data from all configurations, I to V (Table 6.2), and are based on the features selected as optimal by their respective SOFAH-based models from model set I, that is, $\mathrm{FM}_{\mathrm{surf}}$ uses selected features from $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH}}$, and FM_{Δ} from $\mathrm{M}_{\Delta}^{\mathrm{SOFAH}}$, both with the addition of h_1 and h_2. Note that surface temperature T_{surf} is not affected by the different configurations, as varying base and slab thickness has no effect on it. The main purpose of $\mathrm{FM}_{\mathrm{surf}}$ is to predict T_{surf} values to be used in FM_{Δ}, so that the latter does not rely on observations of surface temperature.

The division of data in training and test was 80% and 20% respectively, so model set II was trained on 69888 instances and tested on 17472 instances. All the data from 01/04/2021 to

12/31/2023 was used for inference. In conclusion, model set II and model set I were trained on data from the same time period, 01/04/2019 to 12/31/2020, being the difference between the sets the fact that set I uses data only from configuration I, while set II uses data from all configurations, I-V.

6.2. RESULTS AND DISCUSSION

6.2.1. Data exploratory analysis

Since both models sets I and II were trained on data from 2019 and 2020. Pearson Correlation Matrix of configuration I in this time period is shown in Figure 6.3. Based on the matrix, it can be said that:

- Regarding models that predict T_{air}, relative humidity (RH) shows a high negative correlation with the output, $RH - T_{air}$ (-0.69), and H a moderate correlation, $H - T_{air}$ (0.51);

- Regarding models that predict T_{surf}, none of the features show even moderate correlation with the output, being the highest one $T_{air} - T_{surf}$ (0.37);

- With respect to models that predict ΔT, most of the inputs show low or very low correlation, from $D - \Delta T$ (-0.04) to $P_{atm} - \Delta T$ (0.24), but there is one exception that shows very high positive correlation, $T_{surf} - \Delta T$ (0.96). This very high correlation was one of the reasons for the choice of using T_{surf}^{pred} as input to predict ΔT, even if there is some error accumulation in using predictions as inputs of another prediction.

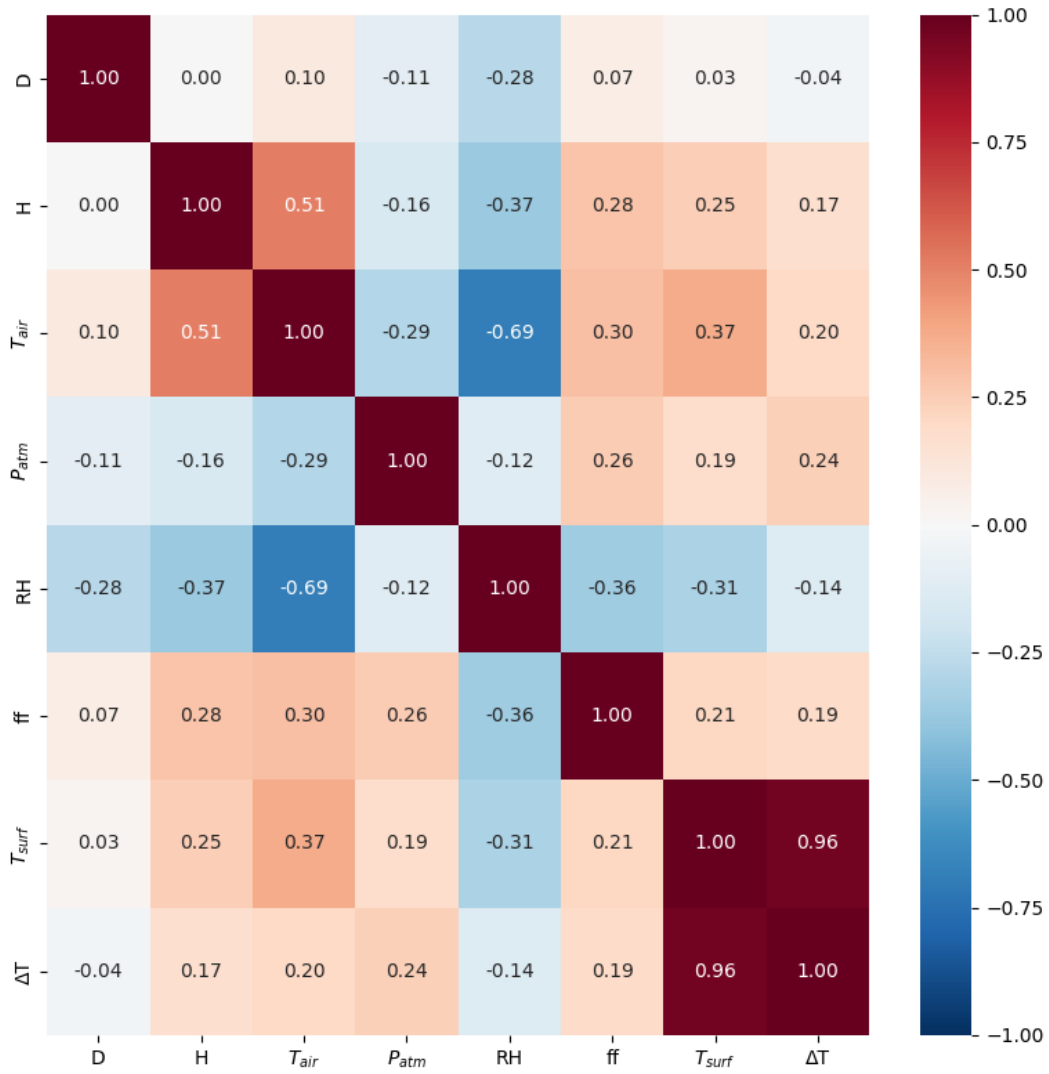


Figure 6.3. Pearson Correlation Matrix of the configuration I dataset

Some statistics were measured for each variable, being them: minimum, maximum, median, mean, SD, skewness, CV, lower quartile (Q1) and upper quartile (Q3). These values are shown only on configuration I in Table 6.6, that is, data used in model set I. Table 6.7 and Table 6.8 presents values for all configurations, I-V, that is, data used in model set II, while also showing values for years 2021, 2022 and 2023, as they were used for inference.

Table 6.6. Variables statistics of training and test data for configuration I

Output	Min	Max	Median	Mean	SD	Skew	CV	Q1	Q3
2019-2020 (Training set)									
D	1.00	366.00	185.00	184.21	105.10	0.00	0.57	93.00	275.00
H	0.00	23.00	12.00	11.50	6.91	0.00	0.60	6.00	17.00
T_{air}	8.70	35.50	21.10	21.65	3.95	0.26	0.18	19.00	24.40
P_{atm}	879.80	896.50	887.60	887.77	2.39	0.19	0.00	886.10	889.30
RH	11.00	97.00	68.00	65.53	21.64	-0.38	0.33	49.00	85.00
ff	0.10	7.30	2.10	2.22	1.11	0.50	0.50	1.40	3.00
T_{surf}	12.83	77.33	27.02	30.36	9.83	1.04	0.32	22.83	36.30
ΔT	-15.92	49.53	-3.80	0.33	9.66	1.07	29.06	-6.64	6.76
2019-2020 (Test set)									
D	1.00	366.00	181.00	183.14	105.06	0.01	0.57	92.00	274.50
H	0.00	23.00	11.00	11.49	6.98	0.01	0.61	5.00	18.00
T_{air}	8.90	35.00	21.10	21.68	4.02	0.29	0.19	19.00	24.40
P_{atm}	880.00	896.50	887.70	887.77	2.43	0.16	0.00	886.10	889.40
RH	10.00	97.00	68.00	65.65	21.85	-0.43	0.33	49.00	85.00
ff	0.10	6.80	2.10	2.23	1.11	0.50	0.50	1.40	3.00
T_{surf}	12.23	70.69	26.97	30.07	9.45	1.06	0.31	22.93	35.34
ΔT	-14.86	36.91	-3.89	-0.02	9.33	1.13	-508.27	-6.69	5.61

Table 6.7. Variables statistics of training and test data for configurations I-V

Output	Min	Max	Median	Mean	SD	Skew	CV	Q1	Q3
2019-2020 (Training set)									
D	1.00	366.00	184.00	183.87	105.09	0.00	0.57	93.00	275.00
H	0.00	23.00	12.00	11.50	6.92	0.00	0.60	5.00	17.00
T_{air}	8.70	35.50	21.10	21.66	3.97	0.27	0.18	19.00	24.40
P_{atm}	879.80	896.50	887.70	887.78	2.40	0.18	0.00	886.10	889.40
RH	10.00	97.00	68.00	65.55	21.68	-0.39	0.33	49.00	85.00
ff	0.10	7.30	2.10	2.22	1.11	0.50	0.50	1.40	3.00
T_{surf}	12.23	77.33	27.04	30.31	9.76	1.04	0.32	22.87	36.08
ΔT	-15.97	49.53	-3.86	0.25	9.34	1.11	37.16	-6.37	6.32

Table 6.8. Variables statistics of training and test data for configurations I-V (continuation)

Output	Min	Max	Median	Mean	SD	Skew	CV	Q1	Q3
2019-2020 (Test set)									
D	1.00	366.00	184.00	184.49	105.05	0.00	0.57	94.00	275.00
H	0.00	23.00	11.00	11.49	6.94	0.00	0.60	6.00	18.00
T_{air}	8.70	35.10	21.10	21.65	3.95	0.27	0.18	19.00	24.30
P_{atm}	879.80	896.50	887.60	887.75	2.40	0.19	0.00	886.10	889.30
RH	10.00	97.00	68.00	65.57	21.67	-0.38	0.33	49.00	85.00
ff	0.10	7.10	2.10	2.20	1.11	0.50	0.50	1.30	3.00
T_{surf}	12.23	68.52	26.87	30.27	9.74	1.04	0.32	22.87	36.11
ΔT	-15.71	37.56	-3.91	0.19	9.32	1.10	48.45	-6.41	6.31
2021									
D	1.00	365.00	183.00	183.00	105.37	0.00	0.58	92.00	274.00
H	0.00	23.00	11.50	11.50	6.92	0.00	0.60	5.75	17.25
T_{air}	8.10	34.80	20.70	21.22	4.02	0.19	0.19	18.60	24.00
P_{atm}	880.20	896.20	887.10	887.21	2.55	0.21	0.00	885.40	888.90
RH	12.00	97.00	67.00	65.54	21.64	-0.35	0.33	48.00	85.00
ff	0.10	7.00	2.00	2.10	1.05	0.52	0.50	1.30	2.80
T_{surf}	12.43	79.78	26.57	30.04	10.17	1.06	0.34	22.33	36.24
ΔT	-14.51	46.71	-4.30	-0.03	9.73	1.13	-375.15	-6.78	6.51
2022									
D	4.00	365.00	184.50	184.50	104.50	0.00	0.57	94.00	275.00
H	0.00	23.00	11.50	11.50	6.92	0.00	0.60	5.75	17.25
T_{air}	5.30	32.70	20.80	21.26	3.96	0.01	0.19	18.60	24.20
P_{atm}	877.30	894.00	887.20	887.18	2.34	-0.19	0.00	885.60	888.80
RH	13.00	96.00	64.00	63.34	20.87	-0.26	0.33	47.00	82.00
ff	0.10	8.10	2.20	2.25	1.04	0.36	0.46	1.50	3.00
T_{surf}	9.70	67.94	26.67	29.97	9.93	1.03	0.33	22.43	35.75
ΔT	-14.68	39.69	-3.52	0.47	9.55	1.05	20.52	-6.42	6.76
2023									
D	1.00	365.00	183.00	183.00	105.37	0.00	0.58	92.00	274.00
H	0.00	23.00	11.50	11.50	6.92	0.00	0.60	5.75	17.25
T_{air}	9.60	34.40	21.50	21.96	4.13	0.17	0.19	19.10	24.90
P_{atm}	879.70	895.80	887.50	887.62	2.32	0.15	0.00	886.00	889.20
RH	17.00	98.00	66.00	64.85	19.70	-0.26	0.30	49.00	83.00
ff	0.10	7.70	1.90	2.01	1.02	0.61	0.51	1.20	2.70
T_{surf}	12.93	74.15	27.39	30.81	10.20	0.98	0.33	23.03	37.20
ΔT	-14.23	46.96	-4.34	0.05	9.83	1.09	187.34	-6.93	6.74

Based on Table 6.6, it can be said that ΔT showed large CV, but that is due to the close-to-zero

value of its mean. Also, training and test sets showed very similar statistics, thus the unseen data is almost in its entirety in the range of the training data, presenting very little extrapolation to be performed by models of set I. Now, based on Table 6.7 and Table 6.8, for model set II, test set data is all in the range of the training data, but the same cannot be affirmed about posterior years data.

In 2021, only T_{air} shows a minimum value of 8.10, lower than the one from training data, 8.70, presenting some degree of extrapolation to be made by the regression model, although it is expected to still provide good accuracy, as the difference from 8.70 to 8.10 is small compared to the mean value of T_{air}. Furthermore, probably just a few points show values of T_{air} lower than 8.70 and this is only one of the input variables. In 2022 and 2023, inference may not present results as accurate as for 2021 data, because they show more variables outside of the training data range. In 2022: T_{air} minimum of 5.30 is lower than 8.70, the minimum from training data; P_{atm} minimum of 877.30 is lower than the minimum of 879.80; and ff maximum of 8.10 is higher than the maximum of7.30. In 2023: P_{atm} minimum of 879.70 is slightly lower than 879.80, the minimum from training data; RH minimum of 98 is lower than the minimum of 97; and ff maximum of 7.70 is higher than the maximum of 7.30.

6.2.2. Model set I

In model set I, the time it took for the FEM models to provide the temperature at the bottom of the concrete slab to evaluate ΔT was of 2 full days. The baseline ML model took 3 hours and 27 minutes in HPO and 2 minutes and 12 seconds in training, thus being much cheaper than the FEM model. When using SOFAH algorithm, the evaluated times were 4 hours and 16 minutes in HPO and 2 minutes and 33 seconds in training, so, there was still a reduction in time compared to the FEM model. For SOFAH2, since it adds a second BO after the one from SOFAH, its total time in HPO was 6 hours and 2 minutes, as the training took the same time as the others.

After performing SOFAH, the inputs of $\mathrm{M}_{\mathrm{air}}^{\mathrm{SOFAH}}$, $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH}}$, and $\mathrm{M}_{\Delta}^{\mathrm{SOFAH}}$ were defined, as shown in Table 6.9. Models $\mathrm{M}_{\mathrm{air}}^{\mathrm{SOFAH2}}$, $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH2}}$, and $\mathrm{M}_{\Delta}^{\mathrm{SOFAH2}}$ use the same inputs. As shown, SOFAH algorithm found that 13 was the optimal number of inputs for T_{air} predictive models and 12 for T_{surf} and ΔT. The addition of T_{surf}^{pred} to the input space of ΔT predictive models was done after

the feature augmentation step of SOFAH algorithm.

Table 6.9. Inputs and outputs of SOFAH-based models of model set I

Model	Inputs	Output
M_{air}^{SOFAH}	$D^{-1}, D^2, logD, H, {P_{atm}}^3, P_{atm} \cdot ff, P_{atm} \cdot RH, RH^{-1}, logRH, RH, ff^{-1}, ff^2, ff^3$	T_{air}
M_{surf}^{SOFAH}	$logD, D \cdot T_{air}, D \cdot ff, D \cdot H, H^3, H \cdot T_{air}, {P_{atm}}^3, {P_{atm}}^{-1}, RH \cdot ff, ff^2, ff^3, {T_{air}}^3$	T_{surf}
M_{Δ}^{SOFAH}	$D^{-1}, D^2, D \cdot T_{air}, D \cdot ff, D \cdot H, H^3, H \cdot P_{atm}, {P_{atm}}^3, RH^2, ff^3, logff, logT_{air}, T_{surf}^{pred}$	ΔT

For the analysis of model set I results, they were divided by response variable, so, initially models M_{air}, M_{air}^{SOFAH} and M_{air}^{SOFAH2} were analyzed. These models were built mostly to evaluate if the proposed methodology and the ANN technique was able to predict air temperature based on current day, current hour, atmospheric pressure, relative humidity and wind speed. T_{air} predictive modeling also served as another opportunity for SOFAH performance evaluation, comparing it to baseline methodology. Some performance metrics are shown in Table 6.10, RMSE, R^2, MAPE and SD of Absolute Percentage Error (SDAPE) on training and test sets.

Table 6.10. Regression evaluation metrics of T_{air} predictive models of model set I

Performance metrics	M_{air}	M_{air}^{SOFAH}	M_{air}^{SOFAH2}
RMSE - Training	0.894	0.910	0.872
RMSE - Test	0.989	0.994	0.980
R^2 - Training	0.949	0.947	0.951
R^2 - Test	0.940	0.939	0.941
MAPE - Training	3.31%	3.37%	3.22%
SDAPE - Training	3.55%	3.67%	3.50%
MAPE - Test	3.66%	3.68%	3.62%
SDAPE - Test	4.06%	4.18%	4.16%

As seen, all three models presented similar very good accuracy, with, on average, R^2 of 0.94 and MAPE of 3.65% on test set. The two-step SOFAH algorithm (SOFAH2) reduced RMSE when compared to the baseline M_{air}, but single-step SOFAH did not, although these differences are almost neglectable. MAPE also was reduced in M_{air}^{SOFAH2}, but the SDAPE increased,

indicating more dispersion on the percentage errors. Model M_{air} predictions of all points are shown in Figure 6.4, plotted against observed values and with +/- 10% error margins. Residuals of test set are shown in Figure 6.5 on a scatter plot and also on a histogram.

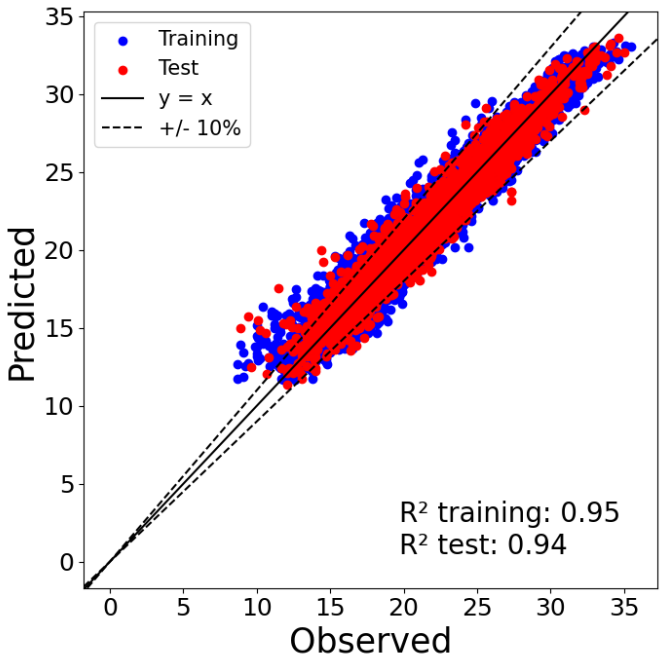


Figure 6.4. Predicted vs. observed values of model M_{air} on training and test sets

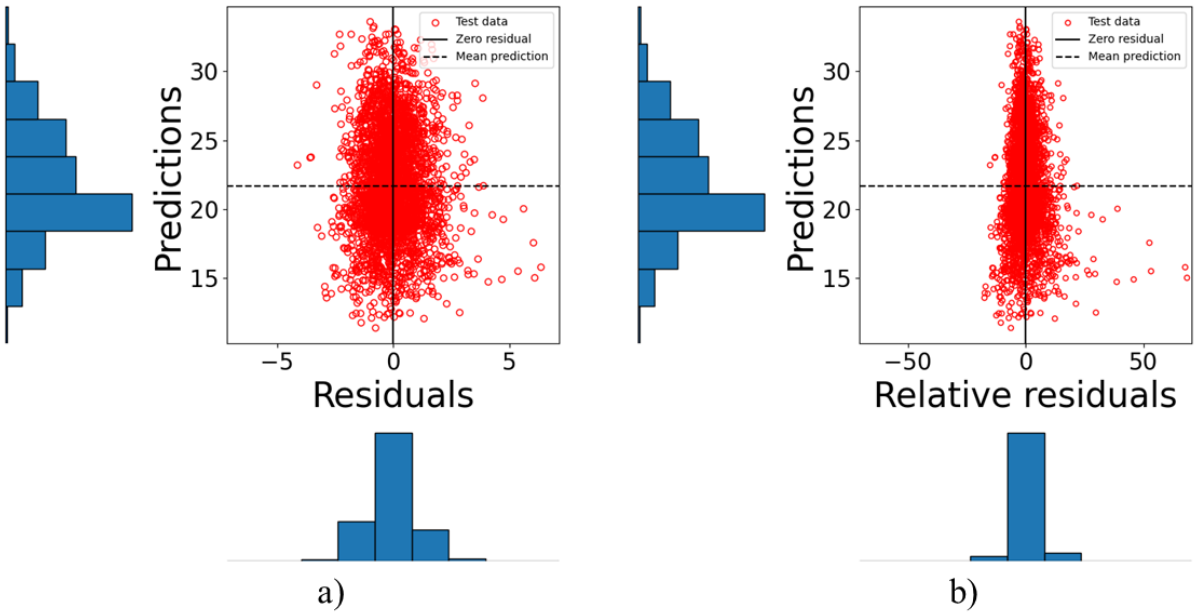


Figure 6.5. Model M_{air} on test set predictions vs.: a) Residuals; b) Relative residuals

As seen in predictions vs. observed plot, the vast majority of test set prediction are inside the +/- 10% error margins, however, it seems that the most inaccurate predictions are all at low values of the response variable, both in training and test. In Figure 6.5 a), it is clear that the largest residuals were positive and at low values of T_{air}. When observing Figure 6.5 b), this behavior is even more noticeable, going past 50% of relative residual. These results indicate that the model did not learn as accurately how to predict low values of T_{air}, but this is generally not a bad result, as both predictions and residuals histograms showed frequency close to that of a Normal distribution.

Permutation Feature Importance (PFI) was evaluated on $N = 30$ permutations of the test set and still using model M_{air}, results are shown in Figure 6.6. As expected, based on the correlation analysis performed previously, RH and H were the most important features for T_{air} prediction, while, surprisingly, D also proved to be relevant to this task. The optimal subset of variables found by SOFAH for T_{air} prediction can be seen in Figure 6.7, with 12 features out of 45.

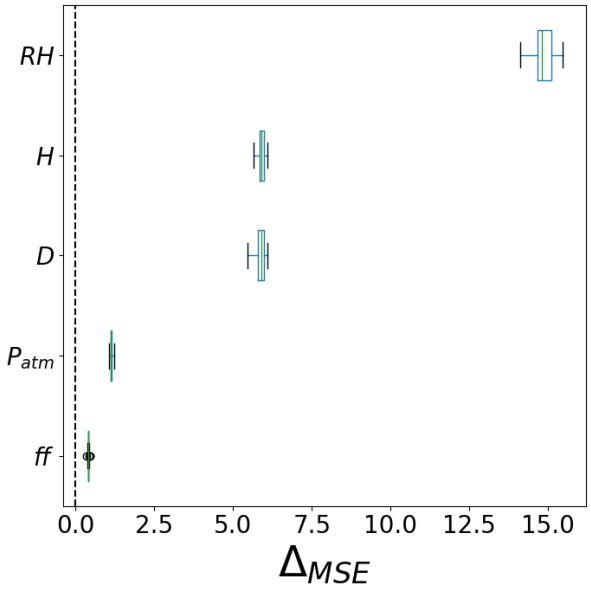


Figure 6.6. PFI results on test set and model M_{air}

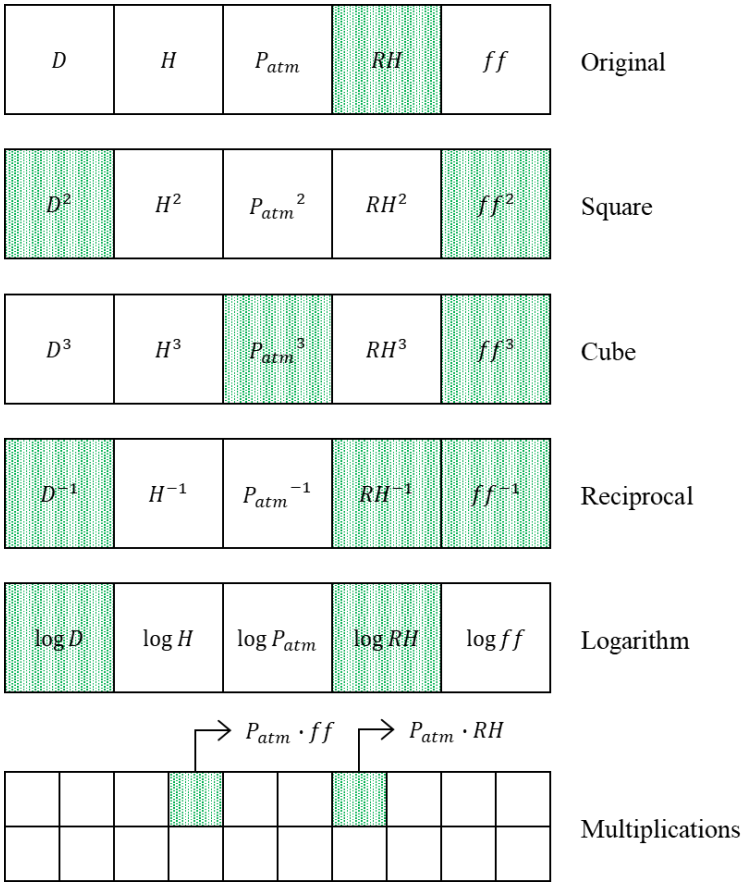


Figure 6.7. Optimal feature subset found by SOFAH for T_{air} prediction

Regression performance metrics were evaluated on T_{surf} predictive models and shown in Table 6.11. The accuracy of these models was outstanding, showing R^2 of nearly 1 and MAPE of less than 1% in all three of them. These results are interesting as no input variable showed high Pearson's correlation with the output. Comparing the models, it can be said that SOFAH clearly improved M_{surf} performance in both training and test sets, reducing RMSE, MAPE, SDAPE and increasing R^2, while SOFAH2 improved even further the already excellent models. To visualize the difference in accuracy between M_{surf} and M_{surf}^{SOFAH2}, two side-by-side comparisons were done in Figure 6.8 and Figure 6.9.

Table 6.11. Regression evaluation metrics of T_{surf} predictive models of model set I

Performance metrics	$\mathbf{M_{surf}}$	$\mathbf{M_{surf}^{SOFAH}}$	$\mathbf{M_{surf}^{SOFAH2}}$
RMSE - Training	0.424	0.366	0.279
RMSE - Test	0.498	0.495	0.412
R^2 - Training	0.998	0.999	0.999
R^2 - Test	0.997	0.997	0.998
MAPE - Training	0.87%	0.73%	0.62%
SDAPE - Training	1.08%	0.89%	0.60%
MAPE - Test	0.95%	0.85%	0.75%
SDAPE - Test	1.28%	1.25%	0.97%

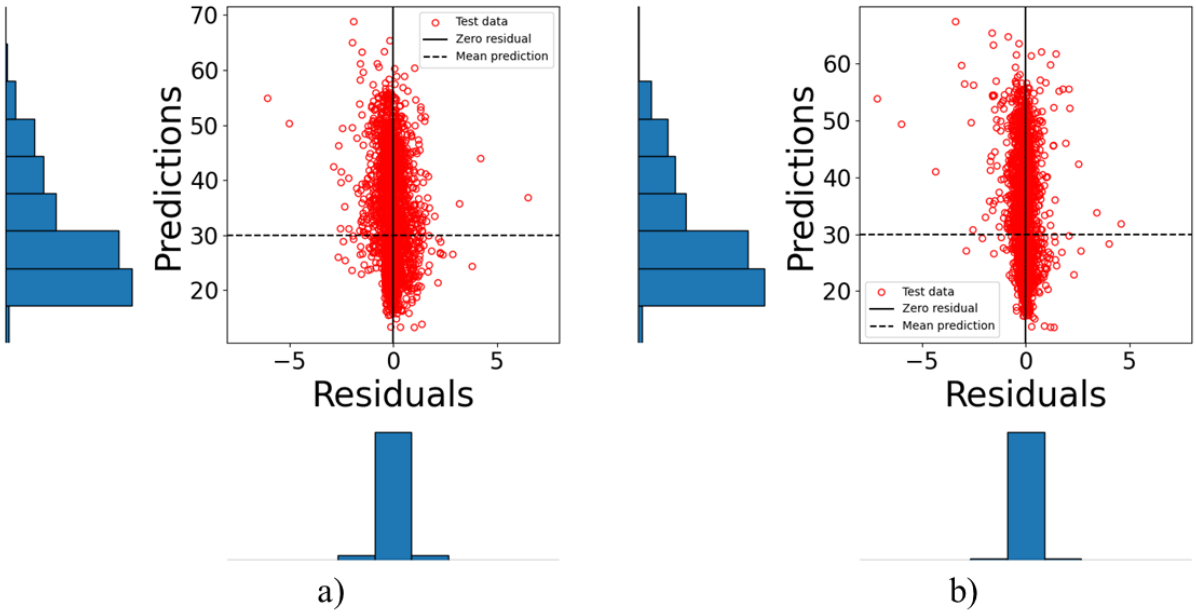


Figure 6.8. Comparison of predictions vs. residuals of: a) M_{surf}; b) M_{surf}^{SOFAH2}

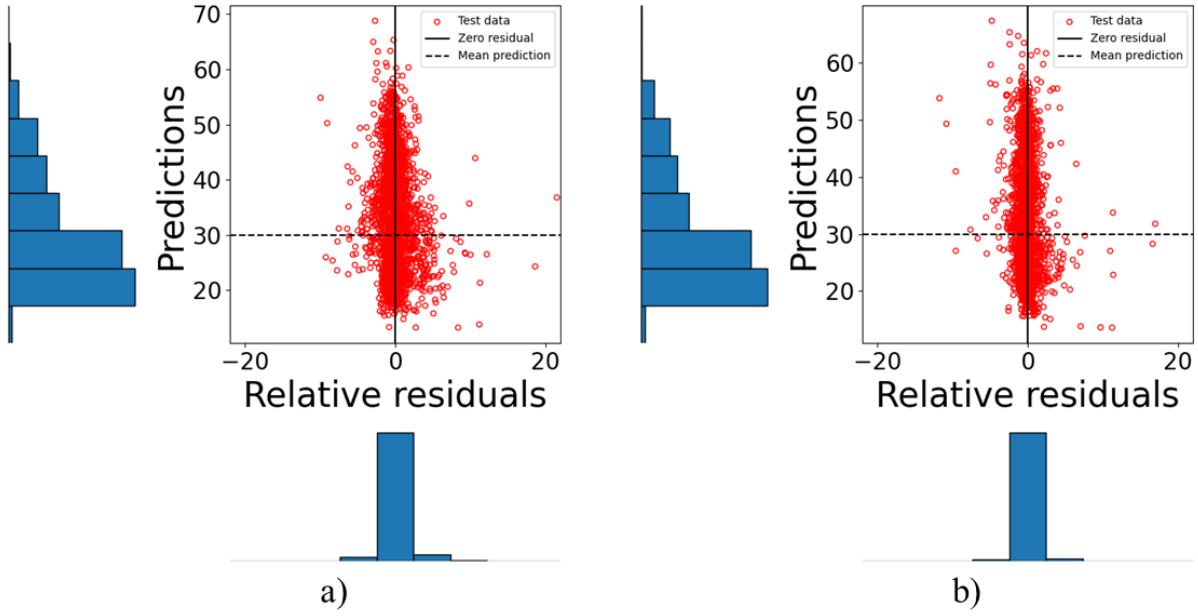


Figure 6.9. Comparison of predictions vs. relative residuals of: a) M_{surf}; b) M_{surf}^{SOFAH2}

It is evident that M_{surf}^{SOFAH2} showed much less dispersed residuals than M_{surf}, thus being generally a more reliable model. Predictions of M_{surf}^{SOFAH2} were plotted against observed values on test set with +/- 10% error margins and shown in Figure 6.10. It is possible to the see how the points align with the diagonal line, indicating highly accurate predictions. This accuracy was one of the motives that led to adding T_{surf}^{pred} as input in predictions of ΔT.

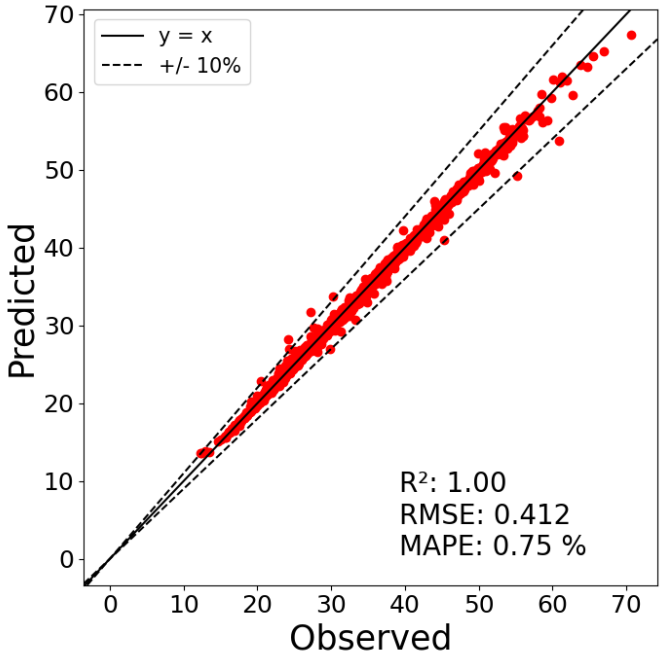


Figure 6.10. Predicted vs. observed values of model M_{surf}^{SOFAH2} on test set

PFI was evaluated with $N = 30$ permutations on test set with model $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH2}}$, thus evaluating feature importance on the optimal inputs selected by SOFAH algorithm (Figure 6.11). As observed, top-5 most important variables consisted of 2 variables of evidently higher importance than the rest, H^3 and $H \cdot T_{air}$, then 2 variables of moderately higher importance, $D \cdot H$ and $D \cdot T_{air}$, and finally, ${T_{air}}^3$. All these 5 variables are composed of only 3 original features, H, T_{air} and D, from which, T_{air} had the highest Pearson's correlation with the output T_{surf}. Also, the optimal subset of variables found by SOFAH in T_{surf} prediction can be seen in Figure 6.12, with 12 features out of 54.

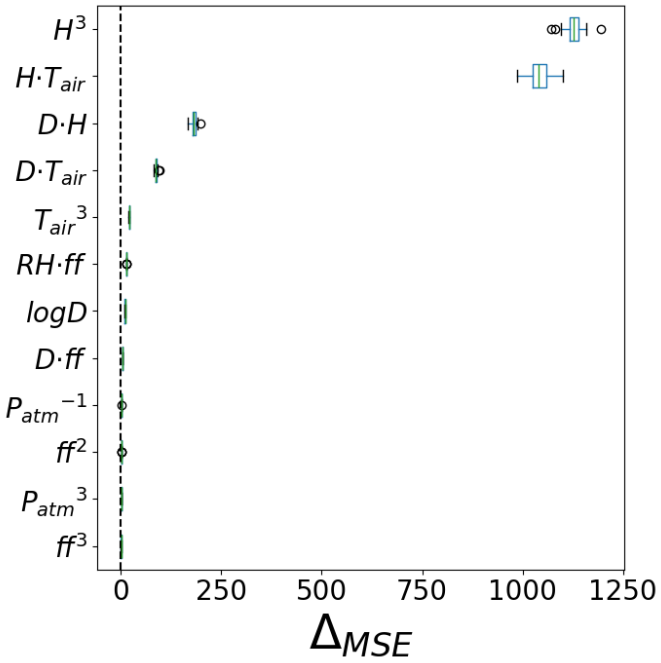


Figure 6.11. PFI results on test set and model $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH2}}$

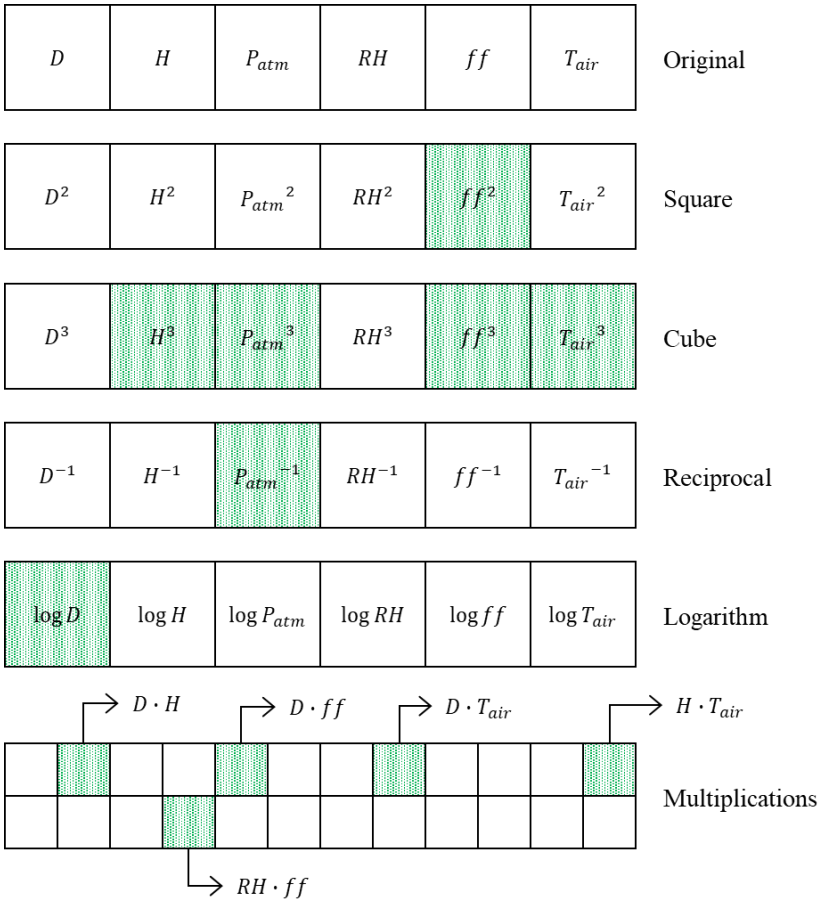


Figure 6.12. Optimal feature subset found by SOFAH for T_{surf} prediction

Regression evaluation metrics were evaluated on T_{Δ} predictive models and shown in Table 6.12. Values of R^2 of 0.98 were found in the three models, but MAPE metrics showed inaccurate results of more than 30%. Despite being widely used, MAPE is biased towards low predictions and cannot be used with zero or close-to-zero values. Based on performed exploratory analysis, the response variable ΔT showed negative minimum, positive maximum and close-to-zero mean. Moreover, very high values of SDAPE indicate that there may be some outliers of extreme relative residuals, as can be seen in Figure 6.13.

Table 6.12. Regression evaluation metrics of ΔT predictive models of model set I

Performance metrics	$\mathbf{M_{\Delta}}$	$\mathbf{M_{\Delta}^{SOFAH}}$	$\mathbf{M_{\Delta}^{SOFAH2}}$
RMSE - Training	0.883	0.968	0.863
RMSE - Test	1.045	1.047	1.006
R^2 - Training	0.992	0.990	0.992
R^2 - Test	0.987	0.987	0.988
MAPE - Training	30.06%	32.03%	28.15%
SDAPE - Training	265.71%	303.79%	245.85%
MAPE - Test	33.66%	36.62%	32.87%
SDAPE - Test	165.19%	220.60%	179.17%

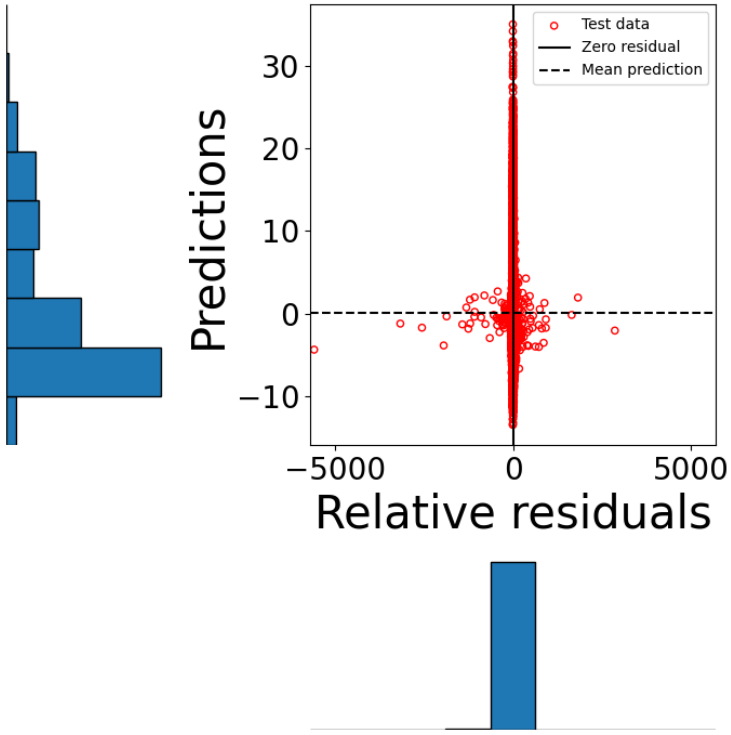


Figure 6.13. Predictions vs. relative residuals for model M_{Δ} on test set

When comparing the models, it can be said that M_{Δ}^{SOFAH} did not improve M_{Δ} performance, providing worse metrics in training and almost no change in test. On the other hand, M_{Δ}^{SOFAH2} did improve the baseline model M_{Δ}, reducing RMSE on test set from 1.045 to 1.006. Predictions are plotted against observed values in Figure 6.14. PFI was evaluated with $N = 30$ permutations on test set with model M_{Δ}^{SOFAH2}, thus evaluating feature importance on the optimal

inputs, but know it is compared with PFI on original feature space, that is, implemented on model M_{Δ}, as shown in Figure 6.15.

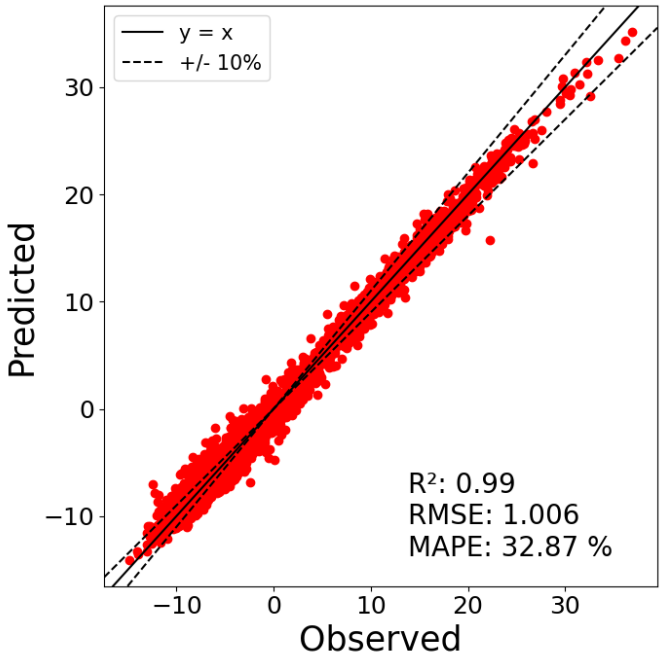


Figure 6.14. Predicted vs. observed values of model M_{Δ}^{SOFAH2} on test set

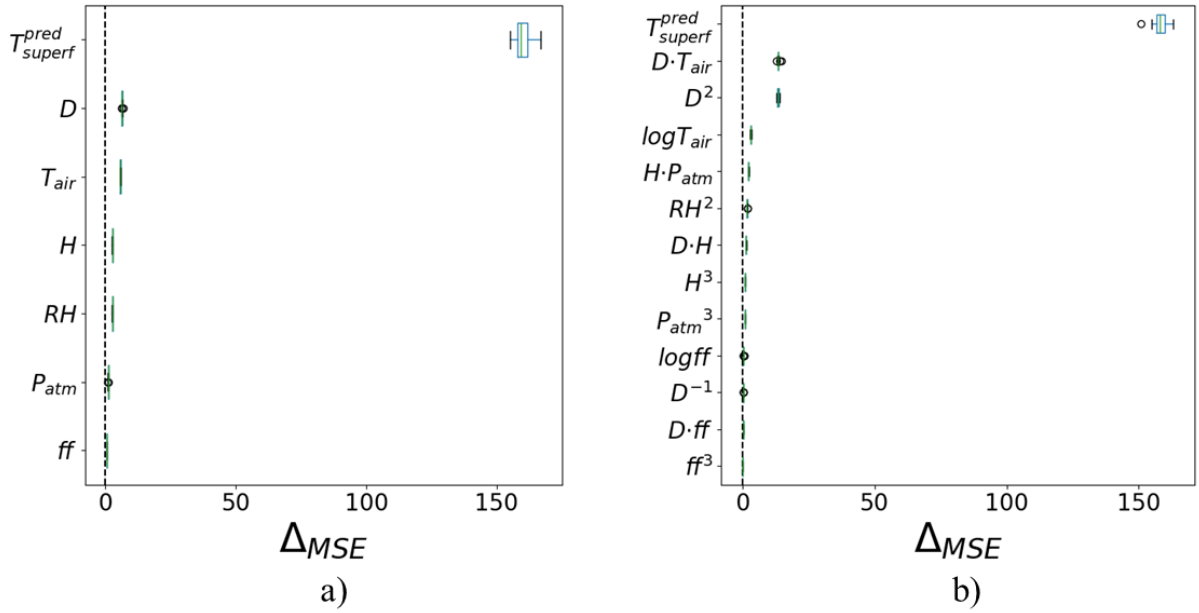


Figure 6.15. Comparison between PFI on test set and models: a) M_{Δ}; b) M_{Δ}^{SOFAH}

As expected, T_{surf}^{pred} showed outstanding importance in both models, as it presented very high correlation with ΔT. Analyzing Figure 6.15 a), features D and T_{air} showed noticeable importance and it is worth noting that, in Figure 6.15 b), second, third and forth most relevant variables consisted of transformations of D and T_{air}. The optimal subset of variables found by SOFAH in ΔT prediction can be seen in Figure 6.16, with 13 features out of 55, T_{surf}^{pred} included.

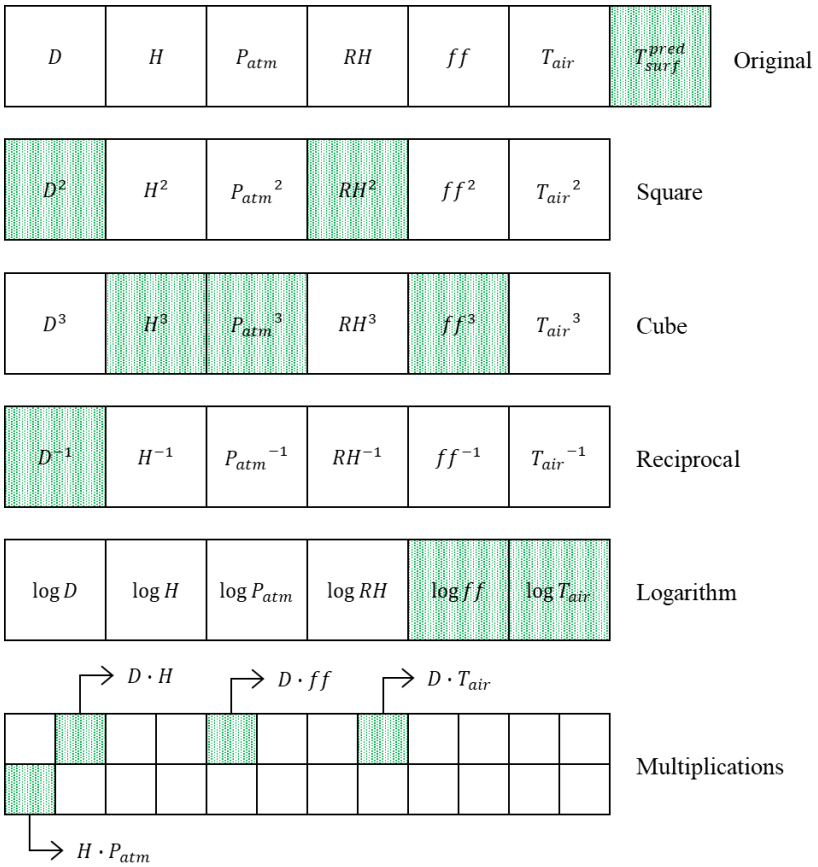


Figure 6.16. Optimal feature subset found by SOFAH for ΔT prediction

Another interesting way to visualize the results is by plotting curves with all test set points over a 1-month period, showing both observed and predicted values. This was done for model $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH2}}$ predicting January 2020 and for $\mathrm{M}_{\Delta}^{\mathrm{SOFAH2}}$ predicting July 2020, respectively shown in Figure 6.17 and Figure 6.18. The daily cycles of increase and decrease in pavement surface temperature and difference in temperature from surface to bottom of the slab can be seen on the plots.

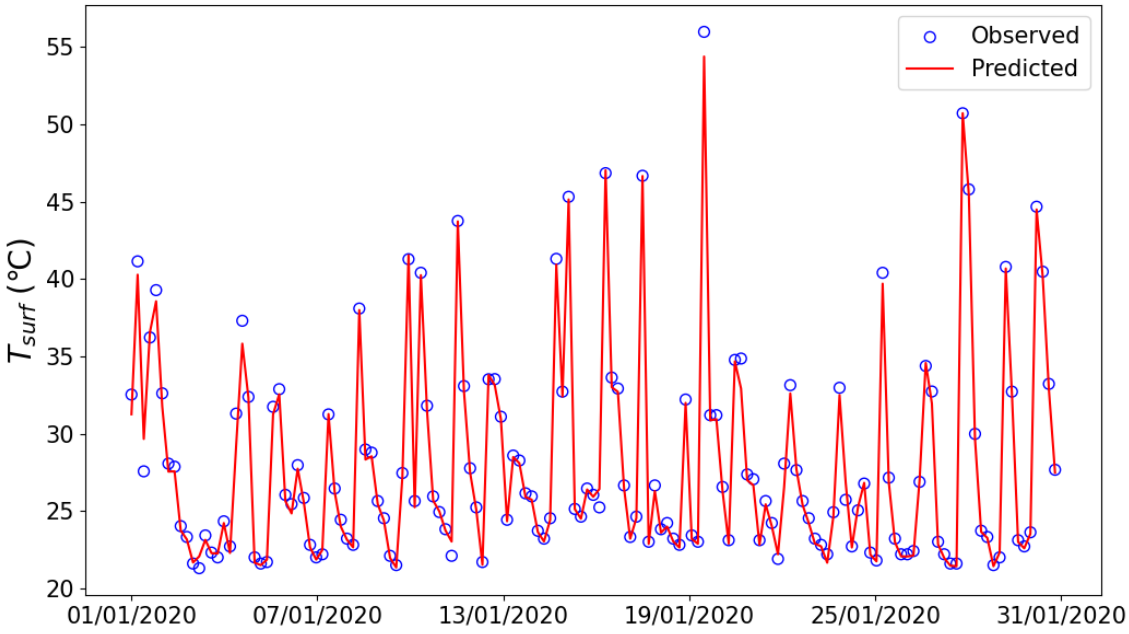


Figure 6.17. Predictions of T_{surf} in January 2020 using model $\mathrm{M}_{\mathrm{surf}}^{\mathrm{SOFAH2}}$

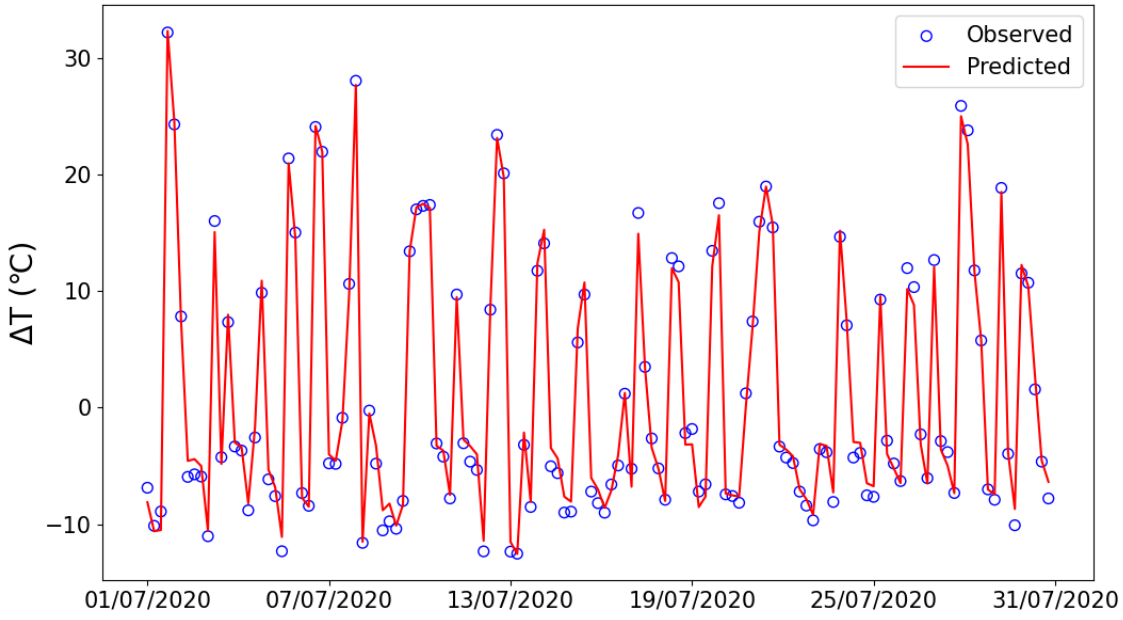


Figure 6.18. Predictions of ΔT in July 2020 using model $\mathrm{M}_{\Delta}^{\mathrm{SOFAH2}}$

6.2.3. Model set II

In model set II, since it used data from 5 configurations, the time it took for the FEM models to provide ΔT was of 10 full days. The amount of data used here was in the tens of thousands, so the ML model took considerable time in HPO and training. The training was performed from 8

to 10 minutes and HPO varied, approximately 7 hours in baseline methodology, 9.5 hours in SOFAH algorithm and 16 hours in SOFAH2. So, even when using datasets of considerable number of instances, the ANNs still were much faster than the FEM models.

The inputs and outputs of model set II, as mentioned, are the same as the ones obtained by the SOFAH algorithm on model set I, with the addition of h_1 and h_2 (Table 6.13). Both h_1 and h_2 were added after the feature augmentation step of SOFAH.

Table 6.13. Inputs and output of each full model (configurations I-V)

Model	Inputs	Output
$\mathrm{FM_{surf}}$	$logD, D \cdot T_{air}, D \cdot ff, D \cdot H, H^3, H \cdot T_{air}, {P_{atm}}^3,$ ${P_{atm}}^{-1}, RH \cdot ff, ff^2, ff^3, {T_{air}}^3, h_1, h_2$	T_{surf}
$\mathrm{FM_{\Delta}}$	$D^{-1}, D^2, D \cdot T_{air}, D \cdot ff, D \cdot H, H^3, H \cdot P_{atm},$ ${P_{atm}}^3, RH^2, ff^3, logff,\ logT_{air}, T_{surf}^{pred}, h_1, h_2$	ΔT

Using data from all five configurations, model set II was created and the predictive accuracy of the models was evaluated on test set, 2021, 2022 and 2023 data separately. Results for model $\mathrm{FM_{surf}}$ are shown in Figure 6.15, from which can be said the accuracy on unseen data from 2019-2020 data was outstanding, achieving R^2 of approximately 1 and MAPE lower than 0.50%. Predictions performed on 2021, 2022 and 2023 data were also outstanding, although not as good as on 2019-2020 data, specially on 2023. This is expected, because of the previously performed analysis on the range of the features that showed some inputs were outside of the training range, specially in 2022 and 2023. Predictions on the mentioned data sets are shown in Figure 6.19.

Table 6.14. Performance metrics of $\mathrm{FM_{surf}}$ predictions on test set, 2021, 2022 and 2023 data

Performance metrics	$\mathrm{FM_{surf}}$ (2019-2020 test set)	$\mathrm{FM_{surf}}$ (2021)	$\mathrm{FM_{surf}}$ (2022)	$\mathrm{FM_{surf}}$ (2023)
MSE	0.056	0.100	0.118	0.555
RMSE	0.236	0.316	0.343	0.745
R^2	0.999	0.999	0.999	0.995
MAPE	0.39%	0.48%	0.52%	1.00%
SDAPE	0.51%	0.73%	1.15%	2.21%

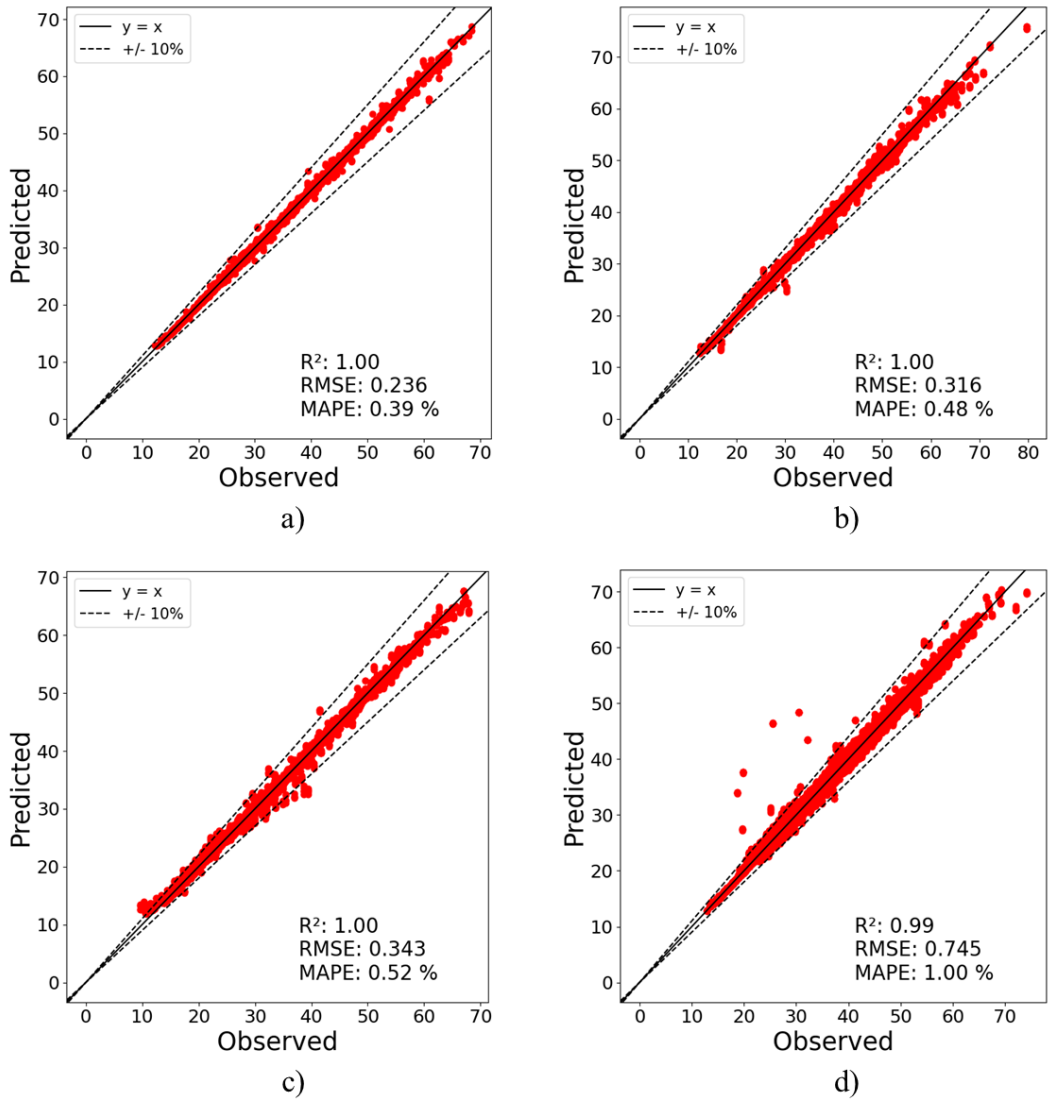


Figure 6.19. Predicted vs. observed values for FM_{surf} on data from: a) 2019-2020 test set; b) 2021; c) 2022; d) 2023

As observed in Figure 6.19, a) provided most accurate predictions, b) and c) showed slightly worse accuracy, and d) had some outliers in regression residuals, with predictions far above the diagonal line and +/- 10% error margins, that is, predicting higher values than observed ones.

The same analyses were done for model FM_{Δ}, as its performance metrics are shown in Table 6.15. In similar way to FM_{surf}, predictions on 2019-2020 test set and 2021 were very good, particularly 2019-2020 and, as expected, 2022 and 2023 showed worse accuracy, although still very accurate, showing R^2 higher than 0.970. MAPE metric provided high values but probably

because of the close-to-zero values of ΔT. Predictions on all these four sets of data are shown in Figure 6.20.

Table 6.15. Performance metrics of FM_{Δ} predictions on test set, 2021, 2022 and 2023 data

Performance metrics	**FM_{Δ} (2019-2020 test set)**	**FM_{Δ} (2021)**	**FM_{Δ} (2022)**	**FM_{Δ} (2023)**
MSE	0.054	1.413	2.430	2.146
RMSE	0.233	1.189	1.559	1.465
R^2	0.999	0.985	0.973	0.978
MAPE	6.31%	37.22%	95.94%	49.71%
SDAPE	43.60%	302.56%	901.45%	436.05%

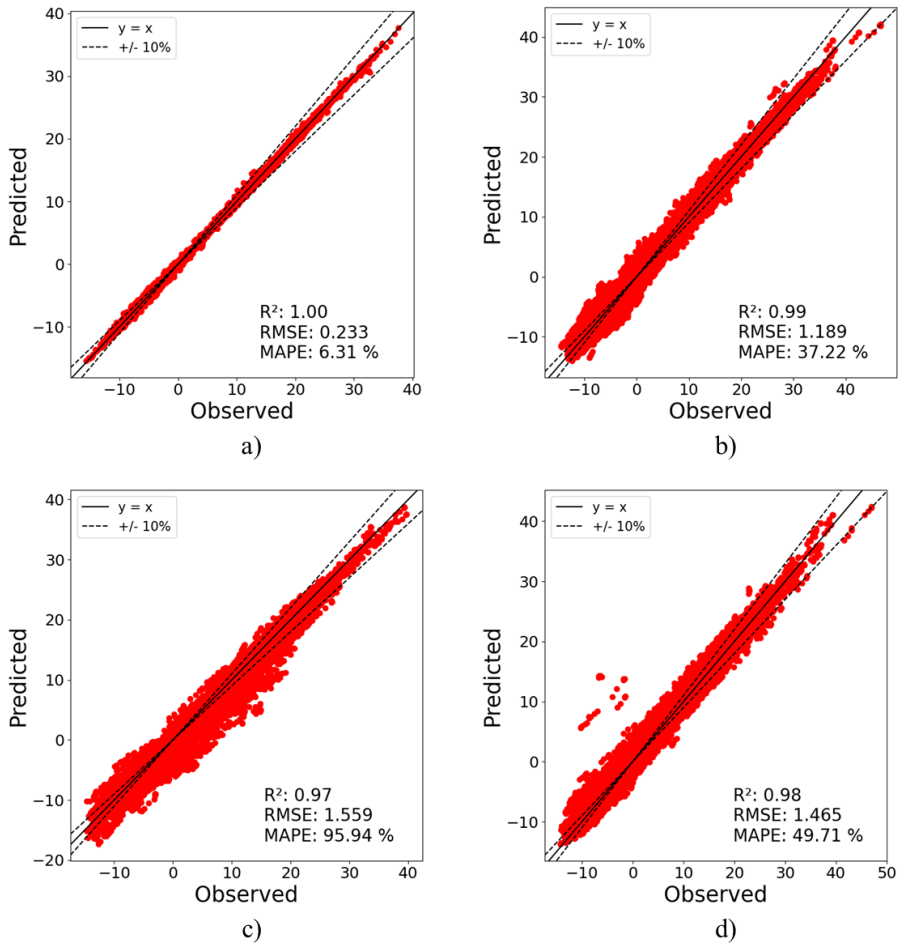


Figure 6.20. Predicted vs. observed values for FM_{Δ} on data from: a) 2019-2020 test set; b) 2021; c) 2022; d) 2023

As seen in Figure 6.20 a), predictions on the test set (2019-2020 data) were extremely accurate, but in b), c) and d), the predictions seem to be more accurate on higher values of ΔT, as the points are between the 10% error margins. In 2022 data, on average, predictions were the least accurate of them all. In 2023 data, it is visually clear that there were some outliers of poor accuracy. Figure 6.21 shows most important features for ΔT predictions, indicating that T_{surf}^{pred} kept the very high importance shown by the models developed only with configuration I data (Figure 6.15). Current day and air temperature also kept considerable importance, but current hour (H) showed increased importance.

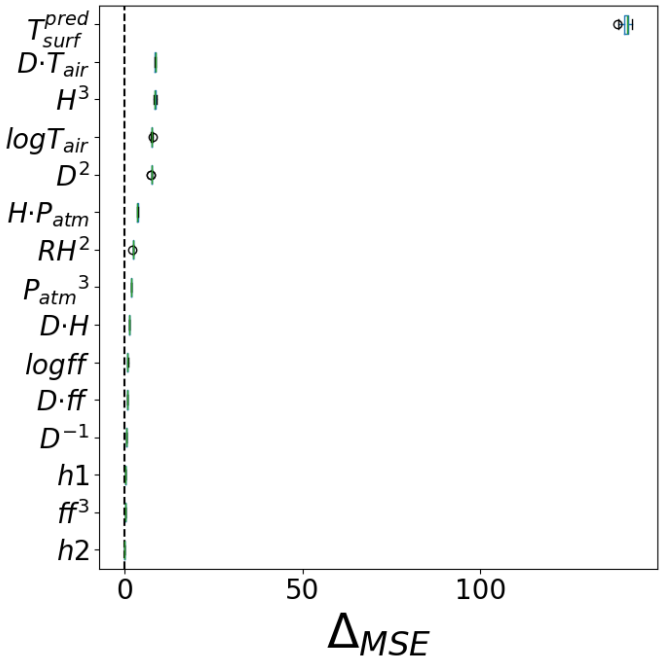


Figure 6.21. PFI on test set and model FM_{Δ}

6.2.4. Uncertainty quantification analysis

By observing the distribution of this model's predictions compared to the distribution of observed values, it is possible to quantify the uncertainty in these models. Both distributions were plotted and shown in Figure 6.22, Figure 6.23, and Figure 6.24 for pavement surface temperature, T_{surf}, respectively in 2021, 2022 and 2023.

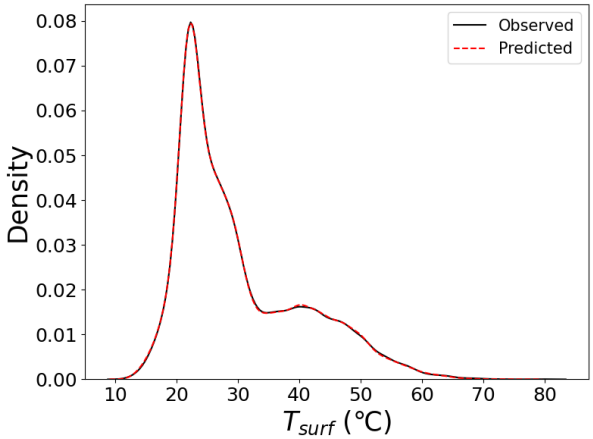


Figure 6.22. Distribution of observed and predicted values of T_{surf} in 2021

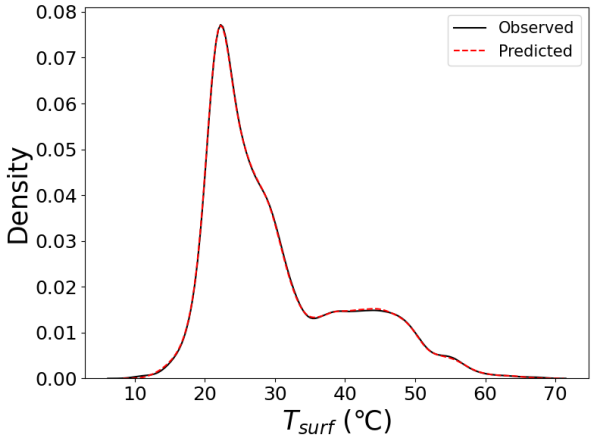


Figure 6.23. Distribution of observed and predicted values of T_{surf} in 2022

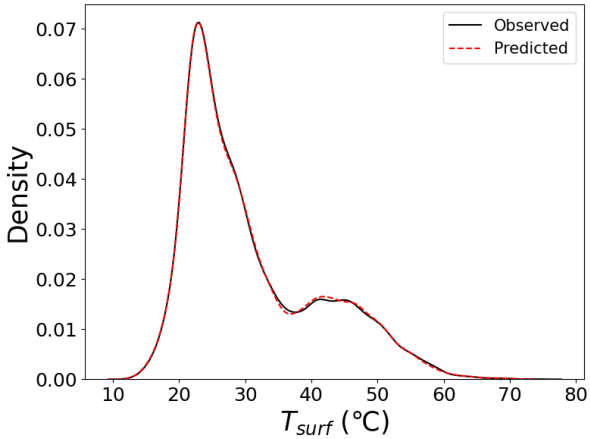


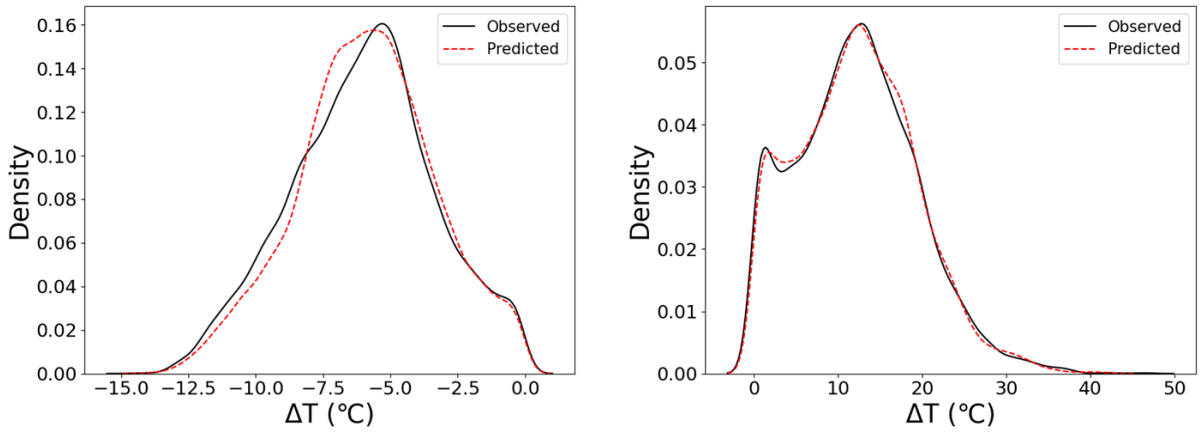
Figure 6.24. Distribution of observed and predicted values of T_{surf} in 2023

As observed, all the three distributions are very similar to each other, in addition, the distribution of predicted values was very close to the distribution of observed values in 2021, 2022 and 2023, highlighting how small the error was in the developed models. Statistics of observations and predictions for 2021, 2022 and 2023 are shown in Table 6.16, complementing the graphical analysis. As observed, almost all the statistics are very similar, for all the three analyzed years, with the only considerable difference in the minimum value in 2022.

Table 6.16. Statistics of observed and predicted T_{surf} in posterior years

Output	2021		2022		2023	
	T_{surf}^{pred}	T_{surf}	T_{surf}^{pred}	T_{surf}	T_{surf}^{pred}	T_{surf}
Min	12.67	12.43	11.87	9.70	12.81	12.93
Max	75.92	79.78	67.66	67.94	70.39	74.15
Median	26.54	26.57	26.61	26.67	27.43	27.39
Mean	30.01	30.04	29.95	29.97	30.81	30.81
SD	10.13	10.17	9.90	9.93	10.19	10.20
Skew	1.04	1.06	1.02	1.03	0.97	0.98
CV	0.34	0.34	0.33	0.33	0.33	0.33
Q1	22.36	22.33	22.43	22.43	23.01	23.03
Q3	36.26	36.24	35.78	35.75	37.37	37.20

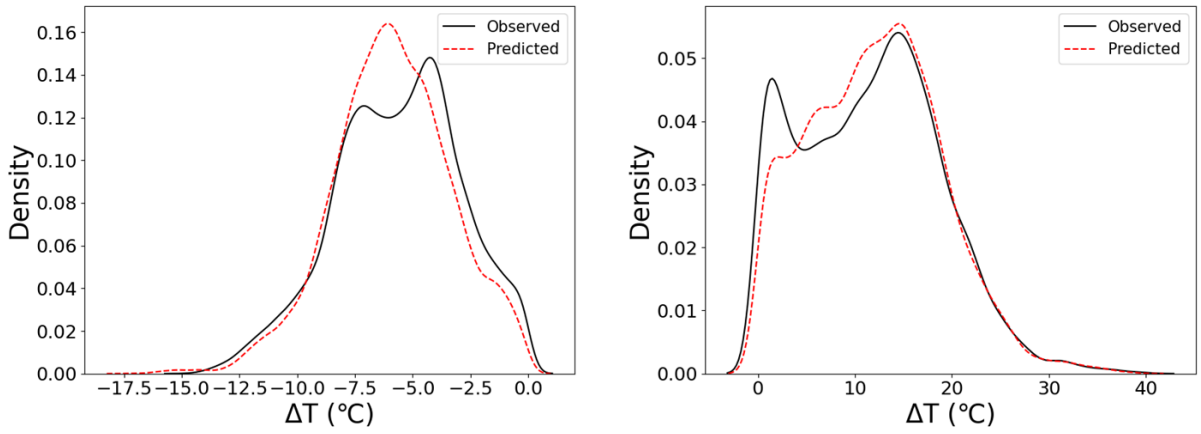
For ΔT predictive models, the same analyses were done, as shown in Figure 6.25, but positive and negative values of ΔT were separated to provide better visualization of days and nights, as days present mostly positive values, since the temperature is higher on the surface than on the bottom of the concrete slab. In nights, it is the opposite, the bottom still has elevated temperature but the air and surface temperatures decrease, thus showing mostly negative values.



a) b)

Figure 6.25. Distribution of observed and predicted values of ΔT in 2021: a) negative; b) positive

As observed, the day shows a bimodal distribution, where the first peak occurs due to the first and last hours of the day, when the temperature at the bottom of the concrete slab is almost the same as at the surface. For most of the day, the second peak occurs, with a positive value of approximately 15 ºC. The night distribution shows only one peak, as ΔT is always negative. During day, observed and predicted distributions are almost identical, as in night it differs a bit. Figure 6.26 and Figure 6.27 show the same comparison, but for 2022 and 2023.



a) b)

Figure 6.26. Distribution of observed and predicted values of ΔT in 2022: a) negative; b) positive

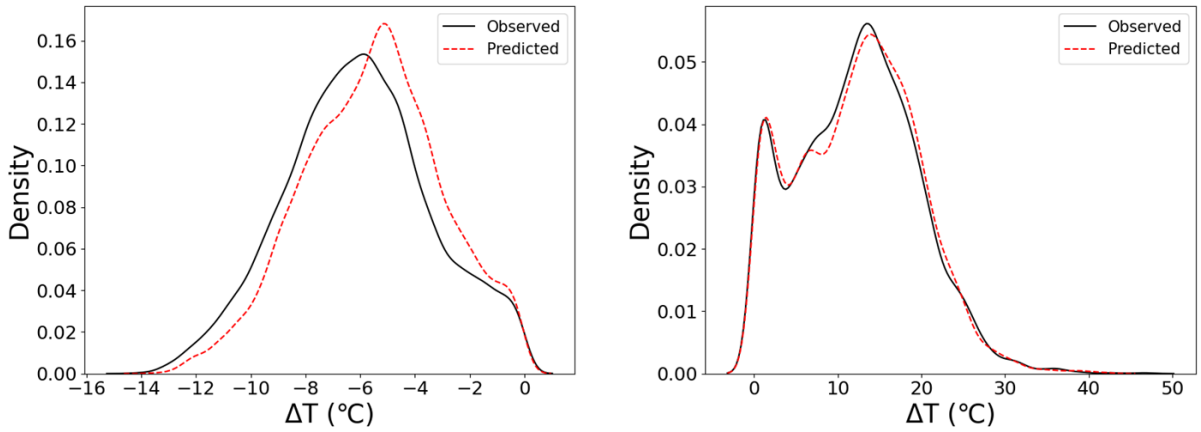


Figure 6.27. Distribution of observed and predicted values of ΔT in 2023: a) negative; b) positive

As expected, in 2022 and 2023 the distribution of predicted values is quite different from the one of observed values in some intervals. The peaks observed in Figure 6.26 a) and b) are clearly some of the intervals of worse predictions. Statistics of observations and predictions are shown in Table 6.17 and Table 6.18, respectively with negative and positive values. Observing the negative values statistics, it is possible to see that they are mostly very similar, with only small differences. Analyzing the positive values statistics, they presented the most differences, highlighting Q1 and mean in 2022. In general, the measured values are similar enough.

Table 6.17. Statistics of observed and predicted negative values of ΔT in posterior years

Output	**2021**		**2022**		**2023**	
	ΔT^{pred}	ΔT	ΔT^{pred}	ΔT	ΔT^{pred}	ΔT
Min	-13.96	-14.51	-17.27	-14.68	-13.66	-14.23
Max	0.00	-0.01	0.00	-0.01	0.00	-0.01
Median	-5.86	-5.89	-5.89	-5.48	-5.44	-6.09
Mean	-5.88	-6.01	-5.89	-5.62	-5.53	-6.07
SD	2.55	2.67	2.55	2.74	2.53	2.67
Skew	-0.08	-0.08	-0.21	-0.24	-0.10	0.04
CV	-0.43	-0.44	-0.43	-0.49	-0.46	-0.44
Q1	-7.53	-7.84	-7.53	-7.51	-7.30	-7.87
Q3	-4.18	-4.26	-4.19	-3.70	-3.81	-4.36

Table 6.18. Statistics of observed and predicted positive values of ΔT in posterior years

Output	**2021**		**2022**		**2023**	
	ΔT^{pred}	ΔT	ΔT^{pred}	ΔT	ΔT^{pred}	ΔT
Min	0.00	0.01	0.00	0.01	0.00	0.01
Max	42.13	46.71	38.79	39.69	42.41	46.96
Median	12.14	12.04	12.04	11.83	12.73	12.48
Mean	12.20	12.14	12.02	11.59	12.34	12.21
SD	7.11	7.21	6.85	7.25	7.27	7.24
Skew	0.37	0.41	0.32	0.31	0.21	0.28
CV	0.58	0.59	0.57	0.63	0.59	0.59
Q1	6.81	6.74	6.67	5.53	6.54	6.65
Q3	16.99	16.87	16.67	16.59	17.48	17.18

7. CONCLUSION

In the present work, GPU-accelerated data-driven models with Artificial Neural Networks (ANN) of vanilla architecture, Multilayer Perceptron (MLP), were created for four different datasets, to predict: residuary resistance of sailing yachts; heating load and cooling load on energy-efficient building; net section capacity of Steel Bolted Angles (SBA) under axial tension; and air temperature, concrete pavement surface temperature (T_{surf}) and difference between temperature on surface and bottom of the concrete slab (ΔT). These models were developed with a baseline methodology and with an algorithm that addresses Hyperparameter Optimization (HPO) and Feature Selection (FS) simultaneously, named SOFAH, Simultaneous Optimization of Feature Augmentation and Hyperparameters.

In general conclusion, it can be stated that:

- PFI algorithm indeed did not handle well highly correlated input features;
- Although a usual metric, MAPE metric proved to be misleading when the response variable presented close-to-zero values, as it showed poor results in very accurate models;
- Baseline methodology proved to be very efficient, providing greatly accurate ML models;
- The chosen details of the baseline methodology performed well: optimization of hyperparameters with BO, parallelization strategies, convergence criterium for BO, usage of AMSGrad optimizer in training, inference with ensemble behavior, choice of regularization techniques, etc;
- Insights on important features on the analyzed engineering problems were acquired;
- Even if not providing a proper equation for prediction, the accuracy of the ANN models was outstanding;
- Machine Learning techniques and data-driven modeling proved to be powerful tools for predicting variables in different engineering-related problems.

7.1. UC IRVINE REPOSITORY DATASETS

Yacht Hydrodynamics (YAH) and Energy Efficiency (ENEFF) datasets were used for performance evaluation of the presented baseline methodology and SOFAH algorithm, as both datasets are publicly available in the widely known UC Irvine repository (KELLY; LONGJOHN; NOTTINGHAM, n.d.). Because of that, many results are available in published researches to compare with the ones obtained in the present work. So, it can be concluded that:

- Very Accurate ANN models were created for the given datasets, achieving, in test set, R^2 of at least 0.99;
- When comparing to other authors results, the presented methodology and SOFAH algorithm showed better metrics and accuracy overall;
- Based on the comparison, the two main methodologies of ML modeling implemented in the present work are competitive.

7.2. COLD-FORMED STEEL BOLTED ANGLES NET SECTION CAPACITY

The investigation of SBA net section resistance was motivated by the debilitating effects of shear lag and eccentricity of connection that make the evaluation of this property not trivial. Furthermore, there is no consensus on its mathematical formulation on standards equations across the world nor on equations proposed by other researchers. These elements are commonly connected by only one leg in transmission towers and communication towers. Based on the developed research, it can be affirmed that:

- PFI evaluation showed that, among the analyzed features, thickness and width of the connected leg were the most important variables when calculating net section capacity of cold-formed steel bolted angles under axial tension in the developed models;
- The addition of numerical data to the experimental training data did not improve accuracy in all cases, but provided the best performing non-SOFAH model, $M_{Exp-Num/5}$, showing, on the test set, R^2 of 0.98, MAPE of 5.70% and RMSE of 5.43;
- Models with reduced number of inputs showed satisfactory accuracy, as even the ones with only 3 inputs had good values of regression evaluation metrics;

- Even when compared with some of the most known standards in use across the world and proposed equations by relevant works on the theme, ML modeling proved to be very efficient and accurate;

- The incorporation of predictions from standards' equations to the training set created more conservative ML models that perform better in a wider domain;

- Specifically for the test set analyzed, AISI equation was much more accurate than Eurocode one;

- AISI equation tends to give highly conservative results for angles with high values of $\bar{x}/L$.

7.3. TEMPERATURE VARIATIONS IN CONCRETE PAVEMENT

Complex phenomena are involved in heat transfer on concrete pavement, affected by solar radiation, conduction between layers and convection. As there is a growing trend in Brazil of using more rigid pavements, predicting accurately the surface temperature and differential temperature through the system's layers is fundamental to design these pavements. Regarding the theme, it can be said that:

- The trained ANNs predicted accurately T_{air}, T_{surf} and ΔT, either when considering a single configuration of concrete pavement or when varying base and concrete slab thicknesses;

- Concrete slab and base thickness of the pavement system were successfully incorporated as features in the ML models;

- Considering all the models, the performance metrics achieved were of at least 0.94 for R^2 on test set, thus representing highly accurate models;

- For air temperature, T_{air}, prediction, relative humidity showed very high importance, and current day and current hour moderate importance;

- In T_{surf} predictive models, it was clear that current hour, air temperature and current day were the most important variables;

- For ΔT predictions, T_{surf}^{pred} , that is, predicted surface temperature, was fundamental to the high accuracy achieved, but also, current day and air temperature showed noticeable importance;

- Models trained on 2019 and 2020 data were able to predict quite accurately T_{surf} and ΔT on 2021, 2022 and 2023 data, even if performing some extrapolation;

- The distribution of predictions in posterior years was analyzed and compared with the distribution of observed values, measuring the uncertainty present in the developed models and their predictions;

- The developed MLP proved to be great surrogate models for the FEM models developed on ABAQUS, as they had much lower computational cost.

7.4. SIMULTANEOUS OPTIMIZATION OF FEATURE AUGMENTATION AND HYPERPARAMETERS

The implemented SOFAH algorithm worked as a second methodology in the present work, where additional information was introduced into the original feature set by performing square, cube, reciprocal, and logarithm transforms, and multiplications of all combinations of two features. Optimal, or close to optimal, subsets of features were found by the algorithm automatically, based on Bayesian optimization. About this topic, it can be stated that:

- The algorithm was successfully implemented, applying BO to simultaneous HPO and FS;

- SOFAH improved almost all of the already very accurate baseline models, while adding information to them, but with additional computational cost;

- The best increase in performance achieved by SOFAH was an improvement of 12% in RMSE of a SBA model;

- Most of the times, the two-step methodology, SOFAH2, was required to show any improvement, resulting in an even bigger increase in computational cost;

- The time taken by the algorithm varied a lot and proved to be highly dependent on the

number of iterations performed in Bayesian Optimization, thus depending on the achievement of the convergence criterion;

- If the objective is to obtain the most accurate model, SOFAH is a viable and competitive algorithm, but if a very accurate model suffices, it added computational cost with little improvement in performance.

7.5. SUGGESTIONS FOR FUTURE WORKS

In future works, it is suggested that:

- More data is acquired regarding net section failure in SBA to form a larger database, aiming to train ML models in wider ranges of variables and obtain better accuracy;

- The numerical database of ABAQUS simulations on concrete pavement is expanded through new simulations, including temperatures on varying depths of the concrete slab and, possibly, resulting in multi-output regression models that predict all of them simultaneously;

- SOFAH algorithm is further developed, testing other optimization algorithms, like genetic algorithms;

- More feature importance techniques are tested, like Shapley additive explanations;

- More uncertainty quantification analyses are performed;

- Some principles of transfer learning and physics-informed neural networks are implemented in the study.

REFERENCES

AMERICAN IRON AND STEEL INSTITUTE (AISI). **AISI S100-16, North American Specifications for the Design of Cold-Formed Steel Structural Members**, 2016.

ARNALDO, Ignacio; O'REILLY, Una-May; VEERAMACHANENI, Kalyan. **Building predictive models via feature synthesis**. In: Proceedings of the 2015 annual conference on genetic and evolutionary computation. 2015. p. 983-990.

ASSOCIAÇÃO BRASILEIRA DE CIMENTO PORTLAND. **Pavimento de concreto avança nas rodovias federais.** Oct. 2024. Available at: https://abcp.org.br/pavimento-de-concreto-avanca-nas-rodovias-federais/. Accessed in: 5 Feb. 2025.

AUTODESK SUPPORT. **Ecotect Analysis Discontinuation FAQ.** Oct. 2024. Available at: https://www.autodesk.com/support/technical/article/caas/sfdcarticles/sfdcarticles/Ecotect-Analysis-Discontinuation-FAQ.html. Accessed in 9 Feb. 2025.

BARBER, E. S. **Calculation of maximum pavement temperatures from weather reports**, Highway Research Board Bulletin 168, National Research Council, Washington, D.C., 1957, pp. 1–8.

BEHZADI-SOFIANI, B.; GARDNER, L.; WADEE, M. A.; DINIS, P. B.; CAMOTIM. D. **Behaviour and design of fixed-ended steel equal-leg angle section columns**. Journal of Constructional Steel Research, 182, 106649, 2021.

BELETE, D. M.; HUCHAIAH, M. D. **Grid search in hyperparameter optimization of machine learning models for prediction of HIV/AIDS test results**. International Journal of Computers and Applications, v. 44, n. 9, p. 875-886, 2022.

BETRÒ, Bruno. **Bayesian methods in global optimization**. Journal of Global Optimization, v. 1, p. 1-14, 1991.

BISHOP, C. M. **Neural networks for pattern recognition.** Oxford University Press, 1995.

BISHOP, C. M. **Pattern Recognition and Machine Learning.** Springer, 2006.

BISHOP, C. M.; BISHOP, H. **Deep learning: Foundations and concepts**. Springer Nature, 2023.

BOUTHILLIER, Xavier; VAROQUAUX, Gaël. **Survey of machine-learning experimental methods at NeurIPS2019 and ICLR2020**. 2020. Tese de Doutorado. Inria Saclay Ile de France.

BRADBURY, R. D. **Reinforced concrete pavements**. 1938.

BREIMAN, L. **Random Forests**. Machine Learning 45, 3-32, 2001.

BROCHU, E.; CORA, V. M.; FREITAS, N. **A tutorial on Bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning**. arXiv preprint arXiv:1012.2599, 2010.

BURDICK, Arlan. **Strategy Guideline: Accurate heating and cooling load calculations**. National Renewable Energy Lab. (NREL), Golden, CO (United States), 2011.

CAI, J.; LUO, J.; WANG, S.; YANG, S. **Feature selection in machine learning: A new perspective**. Neurocomputing, v. 300, p. 70-79, 2018.

CHAIBI, M.; TARIK, L.; BERRADA, M.; EL HMAIDI, A. **Machine learning models based on random forest feature selection and Bayesian optimization for predicting daily global solar radiation**. International Journal of Renewable Energy Development, v. 11, n. 1, p. 309, 2022.

CHEN, Jiaqi; WANG, Hao; XIE, Pengyu. **Pavement temperature prediction: Theoretical models and critical affecting factors**. Applied thermal engineering, v. 158, p. 113755, 2019.

CHESSON JR., E. **Behavior of Large Riveted and Bolted Structural Connections.** Thesis presented to the University of Illinois in partial fulfillment of the requirements for the degree of Doctor of Philosophy. Urbana, 1959.

CHEVALIER, C.; GINSBOURGER, D. **Fast Computation of the Multi-points Expected Improvement with Applications in Batch Selection**. 2012. hal-00732512v2

CHOU, J. S.; TSAI, C. F.; PHAM, A. D.; LU, Y. H. **Machine learning in concrete strength simulations: Multi-nation data analytics**. Construction and Building materials, v. 73, p. 771-780, 2014.

CONFEDERAÇÃO NACIONAL DO TRANSPORTE. **Pesquisa CNT de Rodovias 2023 reforça a importância de maior investimento na malha rodoviária.** Nov. 2023. Available at: https://www.cnt.org.br/agencia-cnt/pesquisa-cnt-de-rodovias-2023-refora-a-importancia-de-maior-investimento-na-malha-rodoviria. Accessed in: 5 Feb. 2025

DEMPSEY, B. J. **A heat transfer model for evaluating frost action and temperature related effects in multilayered pavement systems.** In: Highway Research Record 342, HRB, National Research Council, Washington, D.C., 1970, pp. 39-56.

ELLOBODY, E.; YOUNG, B. **Behavior of cold-formed steel plain angle columns**. Journal of Structural Engineering, 131(3), 457-466, 2005.

EUROPEAN COMMITTEE FOR STANDARDIZATION (CEN). **EN 1993-1-8:2005, Eurocode 3: Design of steel structures - Part 1-8: Design of joints**, CEN, 2005a.

EUROPEAN COMMITTEE FOR STANDARDIZATION (CEN). **EN 1993-1-1:2005, Eurocode 3: Design of steel structures - Part 1-1: General rules and rules for buildings**, CEN, 2005b.

EVANGELISTA JR., F.; ALMEIDA, I. F. **Machine learning RBF-based surrogate models for uncertainty quantification of age and time-dependent fracture mechanics**. Engineering Fracture Mechanics, 258, 108037, 2021.

FAN, Z.; GOU, J.; WANG, C.; CHI, H. **A reduced weighted wang-mendel algorithm using the clustering algorithm to build fuzzy system**. In: 2016 International Conference on Progress in Informatics and Computing (PIC). IEEE, p. 8-12, 2016.

FAN, Z.; GOU, J.; WANG, C.; LUO, W. **Fuzzy model identification based on fuzzy-rule clustering and its application for airfoil noise prediction**. Journal of Intelligent & Fuzzy Systems, v. 33, n. 3, p. 1603-1611, 2017.

FAN, Z.; CHIONG, R.; HU, Z.; DHAKAL, S; LIN, Y. **A two-layer Wang-Mendel fuzzy**

approach for predicting the residuary resistance of sailing yachts. Journal of Intelligent & Fuzzy Systems, v. 36, n. 6, p. 6219-6229, 2019.

FISHER, A.; RUDIN, C.; DOMINICI, F. **All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously**. Journal of Machine Learning Research, vol. 20, n. 177, p. 1-81, 2019.

FLEITAS, I.; BONILLA, J.; BEZERRA, L. M.; MIRAMBELL, E. **Net section resistance in bolted cold-formed steel angles under tension. Journal of Construction Steel Research**, https://doi.org/10.1016/j.jcsr.2019.105841, 2020.

FONTES, B. V. C. **Análise numérica das tensões de cantoneiras de aço formadas a frio, sob tração e conectadas por parafusos**. Dissertação de Mestrado em Estruturas e Construção Civil, Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Brasília, 2020.

FRAZIER, Peter I. **A tutorial on Bayesian optimization**. arXiv preprint arXiv:1807.02811, 2018.

GAL, Y.; GHAHRAMANI, Z. **Dropout as a bayesian approximation: Representing model uncertainty in deep learning**. In: international conference on machine learning. PMLR, 2016. p. 1050-1059.

GAO, Haiping *et al*. **Revolutionizing membrane design using machine learning-bayesian optimization**. Environmental Science & Technology, v. 56, n. 4, p. 2572-2581, 2021.

GEIGER, R. **The climate near the ground**. Harvard University Press, Cambridge, Mass., 1959.

GERRITSMA, J.; ONNINK, R.; VERSLUIS, A. **Yacht Hydrodynamics [Dataset]**. UCI Machine Learning Repository, 1981. https://doi.org/10.24432/C5XG7R.

GHOSH, S.; DOSHI-VELEZ, F. **Model selection in Bayesian neural networks via horseshoe priors**. arXiv preprint arXiv:1705.10388, 2017.

GOODFELLOW, I; BENGIO, Y.; COURVILLE, A. **Deep Learning**. Massachusetts: MIT Press, 2016.

GÖRMEZ, Yasin. **BayesianOptFs: Bayesian Optimization Based Novel Feature Selection Method for Classification Models**. In: 2024 32nd Signal Processing and Communications Applications Conference (SIU). IEEE, 2024. p. 1-4.

GOU, J.; HOU, F.; CHEN, W.; WANG, C.; LUO, W. **Improving Wang–Mendel method performance in fuzzy rules generation using the fuzzy C-means clustering algorithm**. Neurocomputing, v. 151, p. 1293-1304, 2015.

GOU, J.; FAN, Z.; WANG, C.; LUO, W.; CHI, H. **An improved Wang-Mendel method based on the FSFDP clustering algorithm and sample correlation**. Journal of Intelligent & Fuzzy Systems, v. 31, n. 6, p. 2839-2850, 2016.

GUYON, I.; ELISSEEFF, A. **An introduction to variable and feature selection**. Journal of machine learning research, v. 3, n. Mar, p. 1157-1182, 2003.

HAN, S.; HEO, T.; YEUM, C. M.; KIM, K.; CHOI, J.; TIA, M. **Machine Learning Approach to Rapidly Evaluate Curling of Concrete Pavement**. International Journal of Concrete Structures and Materials, v. 18, n. 1, p. 71, 2024.

HE, K.; ZHANG, X.; REN, S.; SUN, J. **Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification**. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 1026-1034, 2015.

HINCHLIFFE, M.; HIDEN, H.; MCKAY, B.; WILLIS, M.; THAM, M.; BARTON, G. **Modelling chemical process systems using a multi-gene genetic programming algorithm**. In: Genetic Programming: Proceedings of the First Annual Conference (late breaking papers). Cambridge: MIT Press, 1996. p. 56-65.

HOLCOMB, R. D.; LABOUBE, R. A.; YU, W. W. **Tensile and bearing capacities of bolted connections**, Second Summary Report, Civil Engineering Study 95-1, Cold- 10 Formed Steel Series, Department of Civil Engineering, Center for Cold-Formed Steel Structures, University of Missouri-Rolla, 1995.

HUANG, Y.; NOJUMI, M. M.; HASHEMIAN, L.; BAYAT, **A. Evaluation of a machine learning approach for temperature prediction in pavement base and subgrade layers in**

Alberta, Canada. Journal of Transportation Engineering, Part B: Pavements, v. 149, n. 1, p. 04022076, 2023.

IOFFE, S.; SZEGEDY C. **Batch normalization: Accelerating deep network training by reducing internal covariate shift**. In: International conference on machine learning, pp. 448-456, 2015.

JIANG, K.; ZHAO, O. **Experimental and numerical studies of stainless steel angle-to-plate connections**. Thin-Walled Structures, 173, 109026, 2022.

JONES, Donald R.; SCHONLAU, Matthias; WELCH, William J. **Efficient global optimization of expensive black-box functions**. Journal of Global optimization, v. 13, p. 455-492, 1998.

KE, K.; XIONG, Y. H.; YAM, M. C.; LAM, A. C.; CHUNG, K. F. **Shear lag effect on ultimate tensile capacity of high strength steel angles**. Journal of Constructional Steel Research, 145, 300-314, 2018.

KELLY, M.; LONGJOHN R.; NOTTINGHAM, K. **The UCI Machine Learning Repository**, https://archive.ics.uci.edu. Acessed in: 9 Feb. 2025.

KANG, Min-Chang; YOO, Doo-Yeol; GUPTA, Rishi. **Machine learning-based prediction for compressive and flexural strengths of steel fiber-reinforced concrete**. Construction and Building Materials, v. 266, p. 121117, 2021.

KINGMA, D. P.; BA, J. **Adam: A method for Stochastic Optimization**. arXiv preprint arXiv:1412.6980, 2014.

KRIZHEVSKY, A.; SUTSKEVER, I.; HINTON, G. E. **Imagenet classification with deep convolutional neural networks**. Advances in neural information processing systems, 25, 2012.

KUSHNER, H. J. **A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise**. Journal of Basic Engineering, 86(1), 97–106, 1964.

LABOUBE, R. A.; YU, W. W. **Tensile and Bearing Capacities of Bolted Connections**. Final Summary Report, University of Missouri-Rolla, Rolla, Missouri, 1995.

LI, X.; LIU, C.; XUE, Y.; XUE, S.; LIAO, S.; ZHOU, Y. **Predicting bearing capacity of angle steel bolted connections using machine learning based on experimental and numerical database**. In Structures (Vol. 67, p. 107014). Elsevier, 2024.

LIMA, João PS; EVANGELISTA JR, F.; SOARES, C. Guedes. **Hyperparameter-optimized multi-fidelity deep neural network model associated with subset simulation for structural reliability analysis**. Reliability Engineering & System Safety, v. 239, p. 109492, 2023.

LOEY, Mohamed; EL-SAPPAGH, Shaker; MIRJALILI, Seyedali. **Bayesian-based optimized deep learning model to detect COVID-19 patients using chest X-ray image data**. Computers in Biology and Medicine, v. 142, p. 105213, 2022.

LOUIZOS, Christos; WELLING, Max. **Structured and efficient variational deep learning with matrix gaussian posteriors**. In: International conference on machine learning. PMLR, 2016. p. 1708-1716.

MAIOLA, Carlos Henrique. **Ligações parafusadas em chapas finas e perfis de aço formados a frio**. Doctoral dissertation. University of São Paulo, São Carlos, 2004.

MALLICK, R. B.; EL-KORCHI, T. **Pavement engineering: principles and practice**. 4th edition, CRC Press, 2023.

MATE, Yash; SOMAI, Neelam. **Hybrid feature selection and Bayesian optimization with machine learning for breast cancer prediction**. In: 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS). IEEE, 2021. p. 612-619.

MCCONAGHY, T. **FFX: Fast, scalable, deterministic symbolic regression technology**. Genetic Programming Theory and Practice IX, p. 235-260, 2011.

MILAD, A. A. *et al*. **Development of a hybrid machine learning model for asphalt pavement temperature prediction**. IEEE Access, v. 9, p. 158041-158056, 2021.

MOČKUS, Jonas. **On Bayesian methods for seeking the extremum**. In: Optimization techniques IFIP technical conference: Novosibirsk, July 1–7, 1974. Springer Berlin Heidelberg, 1975. p. 400-404.

MUKHOTI, J.; STENETORP, P.; GAL, Y. **On the importance of strong baselines in Bayesian deep learning**. arXiv preprint arXiv:1811.09385, 2018.

MUNSE, W. H.; CHESSON JR., E. **Riveted and Bolted Joints: Net Section Design**. Journal of the Structural Division, ASCE, Vol. 89(ST1), pp. 107-126, 1963.

NAIR, V.; HINTON, G. E. **Rectified linear units improve restricted Boltzmann machines**. In Proceedings of the 27th international conference on machine learning (ICML-10), pp. 807-814, 2012.

NATIONAL TRANSPORTATION SAFETY BOARD (NTSB). **Collapse of I-35w Highway Bridge, Minneapolis, Minnesota, August 1, 2007**. Accident Report NTSB/HAR-08/03, PB2008916203, 2008.

NGUYEN, Vu. **Bayesian optimization for accelerating hyper-parameter tuning**. In: 2019 IEEE second international conference on artificial intelligence and knowledge engineering (AIKE). IEEE, 2019. p. 302-305.

NOJUMI, M. M.; HUANG, Y.; HASHEMIAN, L.; BAYAT, A. **Application of machine learning for temperature prediction in a test road in Alberta**. International Journal of Pavement Research and Technology, v. 15, n. 2, p. 303-319, 2022.

PAULA, V.F.; BEZERRA, L.M.; MATIAS, W.T. **Efficiency reduction due to shear lag on bolted cold-formed steel angles**, J. Constr. Steel Res. 64 (2008) 571e583, 2008.

PAULA, J. H. M. **Projeto e Construção de Edifícios de Aço – Perfis Laminados e Soldados**. 9a edição. Departamento de Engenharia Civil e Ambiental, Universidade de Brasília, Publicação Interna, Brasília, 2021.

PFEIL, W.; PFEIL, M. **Estruturas de Aço – Dimensionamento Prático**. 8ª edição. Livros Técnicos e Científicos, Rio de Janeiro, 2009.

PIRYONESI, S. M.; EL-DIRABY, T. **Climate change impact on infrastructure: A machine learning solution for predicting pavement condition index**. Construction and building materials, v. 306, p. 124905, 2021.

QU, S.; GUO, Y.; SUN, Q. **Resistances of high-strength steel equal-leg-angle section columns eccentrically connected by one leg**. Journal of Constructional Steel Research, 191, 107143, 2022.

RASCHKA, Sebastian; PATTERSON, Joshua; NOLET, Corey. **Machine learning in python: Main developments and technology trends in data science, machine learning, and artificial intelligence**. Information, v. 11, n. 4, p. 193, 2020.

RASMUSSEN, C. E.; WILLIAMS, C. K. I. **Gaussian Processes for Machine Learning**. Massachusetts Institute of Technology. MIT Press, 2006.

RAWSON, K. J.; TUPPER, E. C. **Basic ship theory: Hydrostatics and strength**. Butterworth-Heinemann, 2001.

REDDI, S. J.; KALE, S.; KUMAR, S. **On the convergence of Adam and beyond**. arXiv preprint arXiv:1904.09237, 2019.

RUMELHART, D. E.; HINTON, G. E.; WILLIAMS, R. J. **Learning internal representations by error propagation**. In: Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol. 1. Cambridge, MA, USA: MIT Press, pp. 318-362, 1986.

SAROTHI, S. Z.; AHMED, K. S.; KHAN, N. I.; AHMED, A.; NEHDI, M. L. **Predicting bearing capacity of double shear bolted connections using machine learning**. Engineering Structures, vol. 251, 2022a.

SAROTHI, S. Z.; AHMED, K. S.; KHAN, N. I.; AHMED, A.; NEHDI, M. L. **Machine learning-based failure mode identification of double shear bolted connections in structural steel**. Engineering Failure Analysis, 139, 106471, 2022b.

SHAMS, M. Y.; ELSHEWEY, A. M.; EL-KENAWY, E. M.; Ibrahim, A.; TALAAT, F. M.; Tarek, Z. **Water quality prediction using machine learning models based on grid search method**. Multimedia Tools and Applications, v. 83, n. 12, p. 35307-35334, 2024.

SILVA, V. P.; CARVALHO, R. A.; RÊGO, J. H. S.; EVANGELISTA JR., F. **Machine learning-based prediction of the compressive strength of Brazilian concretes: a dual-dataset study**. Materials, v. 16, n. 14, p. 4977, 2023.

SNOEK, J.; LAROCHELLE, H.; ADAMS, R. P. **Practical bayesian optimization of machine learning algorithms**. Advances in neural information processing systems, v. 25, 2012.

SOLAIMANIAN, M.; KENNEDY, T. W. **Predicting maximum pavement surface temperature using maximum air temperature and hourly solar radiation**. Transportation research record, p. 1-11, 1993.

SUN, Shengyang; CHEN, Changyou; CARIN, Lawrence. **Learning structured weight uncertainty in bayesian neural networks**. In: Artificial Intelligence and Statistics. PMLR, 2017. p. 1283-1292.

TABRIZI, Sepideh Emami *et al*. Hourly road pavement surface temperature forecasting using deep learning models. Journal of Hydrology, v. 603, p. 126877, 2021.

TECHE, M. T.; EVANGELISTA JR., F. **Identificação da existência e localização de danificação estrutural utilizando misturas gaussianas e redes neurais artificiais.** Relatório Final de Programa Institucional de Bolsas de Iniciação Científica, Anais do 30º Congresso de Iniciação Científica da Universidade de Brasília e 21º Congresso do Distrito Federal, 2025 [submetido à publicação].

TEH, L. H.; GILBERT, B. P. **Net section tension capacity of cold-reduced sheet steel angle braces bolted at one leg**. Journal of Structural Engineering, vol. 139, n. 3, pp. 328-337, 2013a.

TEH, Lip H.; GILBERT, Benoit P. **Net section tension capacity of cold-reduced sheet steel channel braces bolted at the web**. Journal of Structural Engineering, v. 139, n. 5, p. 740-747, 2013b.

TELLER, L. W.; SUTHERLAND, E. C. **The structural design of concrete pavements**. 1935.

TSANAS, A.; XIFARA, A. **Energy Efficiency [Dataset]**. UCI Machine Learning Repository, 2012a. https://doi.org/10.24432/C51307.

TSANAS, A.; XIFARA, A. **Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools**. Energy and buildings, v. 49, p. 560-567, 2012b.

TURNER, R; ERICKSSON, D.; MCCOURT, M.; KIILI, J.; LAAKSONEN, E.; XU, Z.; GUYON, I. **Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020**. In: NeurIPS 2020 Competition and Demonstration Track. PMLR, 2021. p. 3-26.

VEHREMCAMP, J. E. **Experimental investigation of heat transfer at an air-earth interface.** Transactions, American Geophysical Union, v. 34, no. 1, 1953, pp. 22-29.

VICTORIA, A. H.; MARAGATHAM, G. **Automatic tuning of hyperparameters using Bayesian optimization**. Evolving Systems, v. 12, n. 1, p. 217-223, 2021.

WANG, L. X. **The WM method completed: a flexible fuzzy system approach to data mining**. IEEE Transactions on fuzzy systems, v. 11, n. 6, p. 768-782, 2003.

WANG, D. **Analytical approach to predict temperature profile in a multilayered pavement system based on measured surface temperature data**. Journal of Transportation Engineering, v. 138, n. 5, p. 674-679, 2012.

WANG, D. **Simplified analytical approach to predicting asphalt pavement temperature. Journal of Materials in Civil Engineering**, v. 27, n. 12, p. 04015043, 2015.

WANG, D. **Prediction of time-dependent temperature distribution within the pavement surface layer during FWD testing**. Journal of Transportation Engineering, v. 142, n. 7, p. 06016002, 2016.

WANG, D.; ROESLER, J. R. **One-dimensional rigid pavement temperature prediction using Laplace transformation**. Journal of Transportation Engineering, v. 138, n. 9, p. 1171-1177, 2012.

WANG, D.; ROESLER, J. R. **One-dimensional temperature profile prediction in multi-layered rigid pavement systems using a separation of variables method**. International Journal of Pavement Engineering, v. 15, n. 5, p. 373-382, 2014.

WANG, D.; ROESLER, J. R.; GUO, D. Z. **Analytical approach to predicting temperature fields in multilayered pavement systems. Journal of engineering mechanics**, v. 135, n. 4, p. 334-344, 2009.

WESTERGAARD, H. M. **Computation of stresses in concrete roads**. In: Proceedings, Highway Research Board, p. 90-112, 1925.

WESTERGAARD, H. M. **Stresses in concrete pavements computed by theoretical analysis**. Public roads, 1926.

WESTERGAARD, H. M. **Analysis of stresses in concrete pavements due to variations of temperature**. In: Highway Research Board Proceedings, p. 201-215, 1927.

Williams, C. K. I. and Rasmussen, C. E. **Gaussian processes for regression**. In Touretzky, D. S., Mozer, M. C., and Hasselmo, M. E., editors, Advances in Neural Information Processing Systems 8, pp. 514–520. MIT Press, 1996.

WU, J.; CHEN, X. Y.; ZHANG, H.; XIONG, L. D.; LEI, H.; DENG, S. H. **Hyperparameter optimization for machine learning models based on Bayesian optimization**. Journal of Electronic Science and Technology, v. 17, n. 1, p. 26-40, 2019.

XIAO, L.; LI, Q. Y.; LI H.; REN Q. **Loading capacity prediction and optimization of coldformed steel built-up section columns based on machine learning methods**. Thin-Walled Structures, vol. 180, 2022.

YANG, K.; LIU, L.; WEN, Y. **The impact of Bayesian optimization on feature selection**. Scientific Reports, v. 14, n. 1, p. 3948, 2024.

YIP, A. S. M.; CHENG, J. J. R. **Shear Lag in Bolted Cold-Formed Steel Angles and Channels in Tension**. Structural Engineering Report No. 233, University of Alberta, Edmonton, Canada, 2000.

ŽEGKLITZ, J.; POŠÍK, P. **Symbolic regression algorithms with built-in linear regression**. arXiv preprint arXiv:1701.03641, 2017.

ZEIADA, W.; DABOUS, S. A.; HAMAD, K.; AL-RUZOUQ, R.; KHALIL, M. A. **Machine learning for pavement performance modelling in warm climate regions**. Arabian journal for science and engineering, v. 45, n. 5, p. 4091-4109, 2020.

ZHILINSKAS, A. G. **Single-step Bayesian search method for an extremum of functions of**

a single variable. Cybernetics, v. 11, n. 1, p. 160-166, 1975.