A multi-objective evolutionary approach to discover explainability trade-offs when using linear regression to effectively model the dynamic thermal behaviour of electrical machines.

BANDA, T.M., ZĂVOIANU, A.-C., PETROVSKI, A., WÖCKINGER, D. and BRAMERDORFER, G.

2024

A Multi-Objective Evolutionary Approach to Discover Explainability Tradeoffs when Using Linear Regression to Effectively Model the Dynamic Thermal Behaviour of Electrical Machines

TIWONGE MSULIRA BANDA, ALEXANDRU-CIPRIAN ZĂVOIANU, and ANDREI PETROVSKI, National Subsea Centre, Robert Gordon University, UK
DANIEL WÖCKINGER and GERD BRAMERDORFER, Institute for Electrical Drives and Power Electronics, Johannes Kepler University, Austria

Modelling and controlling heat transfer in rotating electrical machines is very important as it enables the design of assemblies (e.g., motors) that are efficient and durable under multiple operational scenarios. To address the challenge of deriving accurate data-driven estimators of key motor temperatures, we propose a multi-objective strategy for creating Linear Regression (LR) models that integrate optimised synthetic features. The main strength of our approach is that it provides decision makers with a clear overview of the optimal tradeoffs between data collection costs, the expected modelling errors and the overall explainability of the generated thermal models. Moreover, as parsimonious models are required for both microcontroller deployment and domain expert interpretation, our modelling strategy contains a simple but effective step-wise regularisation technique that can be applied to outline domain-relevant mappings between LR variables and thermal profiling capabilities. Results indicate that our approach can generate accurate LR-based dynamic thermal models when training on data associated with a limited set of load points within the safe operating area of the electrical machine under study.

CCS Concepts: • **Computing methodologies** → **Optimization algorithms**; • **Theory of computation** → **Evolutionary algorithms**;

Additional Key Words and Phrases: Data-driven thermal models, electrical machines, linear regression, explainability, problem formalisation, cost vs accuracy, NSGA-II

ACM Reference format:
Tiwonge Msulira Banda, Alexandru-Ciprian Zăvoianu, Andrei Petrovski, Daniel Wöckinger, and Gerd Bramerdorfer. 2024. A Multi-Objective Evolutionary Approach to Discover Explainability Tradeoffs when Using Linear Regression to Effectively Model the Dynamic Thermal Behaviour of Electrical Machines. *ACM Trans. Evol. Learn.* 4, 1, Article 3 (February 2024), 16 pages.
https://doi.org/10.1145/3597618

This work has been supported by the COMET-K2 "Center for Symbiotic Mechatronics" of the Linz Center of Mechatronics (LCM) funded by the Austrian federal government and the federal state of Upper Austria.
Authors' addresses: T. M. Banda, A.-C. Zăvoianu, and A. Petrovski, National Subsea Centre, Robert Gordon University, Aberdeen, UK; emails: {t.banda, c.zavoianu, a.petrovski}@rgu.ac.uk; D. Wöckinger and G. Bramerdorfer, Institute for Electrical Drives and Power Electronics, Johannes Kepler University, Linz, Austria; emails: {Daniel.Woeckinger, Gerd.Bramerdorfer}@jku.at.

2688-3007/2024/02-ART3 $15.00
https://doi.org/10.1145/3597618

1 INTRODUCTION

As the use of data-driven decision-making systems is becoming commonplace today, users are increasingly demanding some form of understanding on how these systems make decisions. This can be particularly important when the goal is to obtain novel scientific insights from observational or simulated data [Roscher et al. 2020]. Roscher et al. [2020] also propose three highly relevant core characteristics that facilitate human understanding and trust of **machine learning (ML)** models: *transparency*, *interpretability*, and *explainability*. While primarily derived from applications that employ fairly complex ML and deep learning techniques to gain scientific knowledge in the natural sciences, these three core characteristics offer a valuable framework for studying **explainable artificial intelligence (XAI)** systems in general as they provide both a welcomed distinction between often intertwined concepts and a way of understanding interactions between these concepts. In the case of ML, Roscher et al. [2020] posits that:

- *Transparency* concerns the different ingredients of a model: structure, individual components, learning algorithm, and how a specific solution is obtained by the algorithm. This aligns closely with the views in Lipton [2018].
- *Interpretability* refers to the ability to "make sense" of a model (and its results) by presenting some of its properties in a way that is understandable to humans. In contrast to transparency, data is always involved when ascertaining interpretability.
- *Explainability* is fairly subjective, often context-dependent, but could be reasoned about using the prior definition from Montavon et al. [2018]: *"An explanation is a collection of features of the interpretable domain, that have contributed for a given example to produce a decision".*

Based on this taxonomy, it is very easy to understand why linear (regression) models are seen as defining the upper (asymptotic) threshold of explainability for ML: their weight values can directly identify attributes that are relevant for prediction making as well as their relative importance. For this reason, linear models have been used to construct understandable proxies of more complex ML approaches like within the **(Local Interpretable Model-Agnostic Explanations) LIME** approach [Ribeiro et al. 2016], where linearity is used to characterise the local neighbourhood of a datum. Given that the good explainability of linear models is often contrasted by their poor performance across numerous modelling scenarios, the main XAI research focus naturally falls on improving the explainability of complex high-performance approaches (e.g, deep neural networks).

Motivated by the characteristics of our real-life application domain, in this study we propose a slightly counter-intuitive approach to developing effective and explainable data-driven models. In essence, we first use synthetic features to augment the modelling power of linear regression models in order to increase their performance on a well-known non-linear task (dynamic thermal modelling). Given that by adding a large set of synthetic features to the interpretable domain, we are likely to impact the *explainability* of the resulting thermal models, the second step of our approach is to apply an iterative model reduction (i.e., regularisation) strategy to reduce the size of the best performing LR models (and thus mitigate the aforementioned *explainability* impact). More importantly, the entire thermal modelling process is governed by a multi-objective optimisation approach that aims to provide decision makers with an overview of the optimal tradeoffs between data collection costs, expected modelling errors, and model *explainability*. The high-level overview of the key components of our approach alongside their interactions is provided in Figure 1.

In order to maximise trust in the generated data-driven thermal models, we have also sought to maximise the *transparency* and *explainability* of the proposed multi-objective approach itself by (i) working with electrical engineers to integrate domain knowledge in the data-driven modelling problem formulation right from the start and (ii) opting for a step-wise formalisation of the

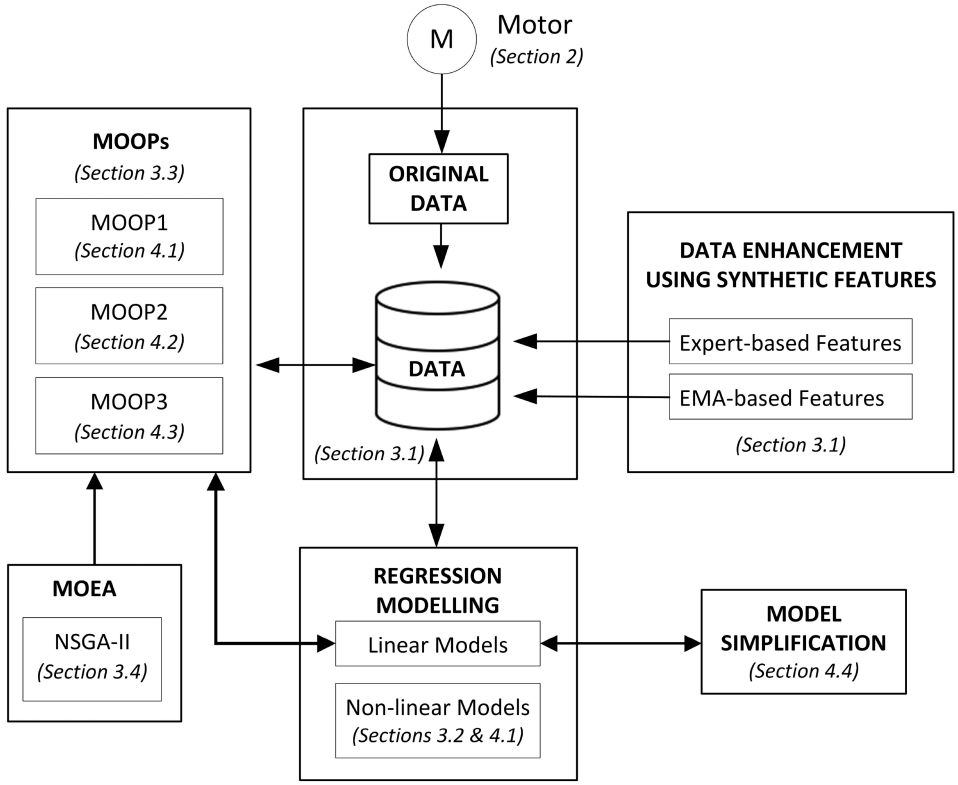

Fig. 1. High-level overview of our data-driven strategy to construct explainable dynamic thermal models.

final multi-objective modelling task that aims to build confidence by incrementally validating key modelling assumptions.

The rest of this article is structured as follows: Section 2 provides a background to thermal modelling for electrical machines and describes the modelling scenario and the requirements that motivate the present work. In Section 3, we describe our multi-objective thermal modelling approach, including data preparation and experimental setup. Section 4 demonstrates the results and provides their interpretation, and finally, Section 5 contains conclusions and an outlook on future work.

2 BACKGROUND TO THERMAL MODELLING OF ELECTRICAL MACHINES

Our industrial case concerns the heat that is produced by electrical machines during their operation. When an electrical machine, e.g. a motor, is running, heat is produced as a result of friction when electrical energy is being converted to mechanical energy. Electrical engineers consider this heat as problematic because, first, it represents losses in efficiency, which may reach up to 25% [Boglietti et al. 2009]; and second, it gradually reduces the lifespan of the electrical machine, and in a worst-case scenario can damage it [Choudhary et al. 2018].

Our case study considers a 3-phase brushless outer rotor permanent magnet synchronous motor, commonly used in low-cost fans. The motor has six key component temperatures that are of interest when wishing to monitor and manage heat (see Figure 2). Domain experts have categorised the components as, high (denoted H), medium (M), and low (L) priority depending on the importance of monitoring their temperature within the general thermal context of the assembly.

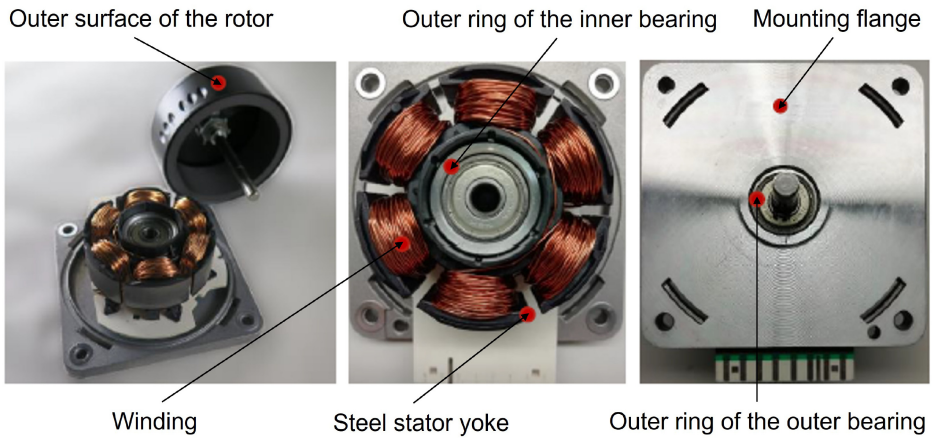

Fig. 2. Electric motor showing the different components (adapted from Wöckinger et al. [2020]).

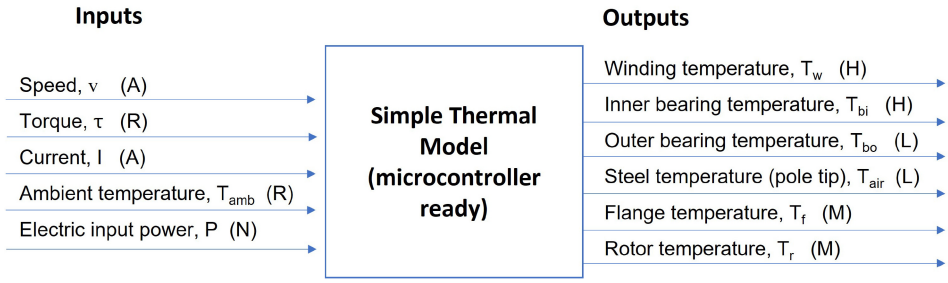

Fig. 3. Summary of modelling requirements with input and output variables.

The high-priority temperatures are for the winding, T_w, and the static ring of the inner ball bearing, T_{bi}. The temperatures of the mounting flange, T_f, and the rotor, T_r are considered of medium priority, whereas the outer ring of the outer ball bearing, T_{bo}, and the steel stator yoke, T_s, are of low priority.

Domain experts have also identified five input variables that are highly relevant for thermal modelling. Depending on the ease and cost of collecting (real-time) sensor data during regular operation, these inputs can be categorised into three groups as follows: A: data always available; R: data rarely available; and N: data never available. The inputs are: rotor speed, v, (A); electric current, I, (A); torque, τ, (R); ambient temperature, T_{amb} (R); and electric power input, P (N). A summary of the modelling requirements is provided in Figure 3.

Traditionally, engineers would turn to the **lumped parameter thermal network (LPTN)** analytical technique to model heat in electrical machines, especially when it comes to accurately modelling the transient thermal processes in the assembly [Boglietti et al. 2009]. However, using LPTN for this kind of motor is known to be challenging [Wöckinger et al. 2020]. While studies like Kirchgässner et al. [2019] and Zăvoianu et al. [2020] have demonstrated the potential of data-driven thermal modelling, researchers also caution that due to the fact that most data-driven models are black-boxes in nature, it is not possible for electrical engineers to obtain particular machine-specific information and thus gain insights from them [Wöckinger et al. 2020]. Therefore, explainability is a key requirement for this data-driven modelling scenario. Further compounding the complexity of the modeling task, data availability restrictions are usually associated with the low-cost applications of these types of motors.

To summarise, our aim is to construct explainable data-driven thermal models that can be used to accurately characterise the real-time dynamic thermal behaviour of electrical machines under different operational scenarios. Using a limited set of data regarding only speed (v) and current (I), the developed thermal models must be able to predict the temperatures of the six above-mentioned output motor components. Furthermore, the models are required to have a simple architecture and be resource efficient in order to facilitate deployment on a microcontroller. For the models to be effective, they should have an average temperature estimation error of less than +/−2°C when the motor is used within its **safe operating area (SOA).**

3 PROPOSED APPROACH TO EXPLAINABLE THERMAL MODELLING

3.1 Data Preprocessing

To enable the creation of models that are applicable under different operational scenarios, domain experts have provided 20 datasets, each containing time series data of temperature profiles that correspond to common usage patterns (load points) of the motor under study. Each dataset contains the two inputs/features (v, I) and six outputs/targets (T_w, T_{bi}, T_f, T_r, T_{bo}, T_s) measured simultaneously at an interval of 2 seconds. Sample sizes for the 20 datasets (marked $DS_{01} \ldots DS_{20}$) range from 571 to 16,201. In total, the 20 datasets contain 240,200 samples (i.e., $\approx$ 133.5 hours worth of testing data). Details of the setup of the test bench, the sensors and the cameras used to collect data are described in Wöckinger et al. [2020] and Wöckinger et al. [2021]. The data itself can be accessed at: https://github.com/czavoianu/TELO_2023.

We used the two provided inputs, speed (v) and current (I), to create two sets of *synthetic features* as follows:

- Based on expert knowledge of electrical machines, torque (τ) is directly proportional to current (I) [Nash 1997] and the total power losses are directly proportional to speed (v) and I [Chalmers and Spooner 1999]. Thus, from a physical point of view, input variables based on several multiplicative combinations of v and I are considered suitable for thermal modelling. We thus created four expert-suggested additional features: v^2, v^3, I^2, and $v \cdot I$. The inclusion of these features is the main channel of incorporating expert knowledge in our modelling approach and arguably improves overall *explainability* by expanding the interpretable domain of our thermal models in a way that is directly aligned with user knowledge and expectations.
- We applied the ***Exponentially Weighted Moving Averages*** **(EMAs)** [Holt 2004] to all the 6 features (2 original + 4 expert-suggested) based on v and I in an effort to smooth random fluctuations in the time series data and complement data samples with information regarding trends. All EMA features were calculated using the formula in Equation (1):

$$\mathrm{EMA}_{\alpha,t}(r) = \alpha \times r_t + (1 - \alpha) \times \mathrm{EMA}_{\alpha,t-1}(r) \tag{1}$$

  where, α is the weight, t is the current period, and r_t is the value of the time series r in the current period. A key aspect when using EMA is to decide how much weight to give to older observations. We initially used weights of 0.001, 0.005, and 0.04 to capture long-, medium-, and short-term trends in the data. A further 18 synthetic inputs were thus created using EMAs, and in total, each of the 20 datasets contained 24 features.

It is important to note that the usage of synthetic EMA features is both a necessity for capturing temporal aspects and a common practice for time series modelling in other fields (e.g., financial and economic modelling). However, the particular number and choice of EMA weights was subjective and largely informed by the authors' modelling experience. As such, this can be seen as negatively impacting the (design) *transparency* of our thermal models.

Table 1. Performance Comparison on Test Data for Different Regression Modelling Techniques

Qual. indicator	RF		KNN		ANN		LR	
	T_w	T_{bi}	T_w	T_{bi}	T_w	T_{bi}	T_w	T_{bi}
MSE	**0.023**	**0.016**	0.076	0.055	0.102	0.080	1.023	0.876
MAE	**0.040**	**0.034**	0.107	0.091	0.202	0.178	0.778	0.724
R^2	**0.999**	**0.999**	**0.999**	0.998	**0.999**	0.997	0.991	0.974

The best result for each (component temperature, quality indicator) pair is highlighted in bold font.

3.2 Preliminary Modelling Insights

To determine the effectiveness of the provided datasets in modelling the target temperatures, we carried out preliminary modelling of the high priority temperatures (T_w and T_{bi}). We combined all the 20 datasets into a single dataset, shuffled it, and, using a simple train-test split, randomly partitioned it into a training set containing 90% of the samples and a test set with the remaining 10% samples. Then, we trained four learning algorithms, **Linear Regression (LR)** [Kutner et al. 2005], **Random Forest (RF)** [Breiman 2001], **K-Nearest Neighbour** (KNN) [Cover and Hart 1967] and a shallow Artificial Neural Network (ANN) [Haykin 1999]. We identified the best parameters for RF (i.e. number of features and maximum depth) and KNN (leaf size and the number of neighbours) using GridSearchCV with 10-fold cross validation available in Scikit-learn [Pedregosa et al. 2011]. For ANN, we used the RandomizedSearchVC with 3-fold cross validation (also available in Scikit-learn) to identify the best configuration for hidden layer sizes, activation, alpha and learning rate.

Results from the preliminary modelling are presented in Table 1 and are largely consistent with previous findings in the sense that non-linear techniques are more accurate in predicting the target temperatures when compared to LR [Zăvoianu et al. 2020]. However, Linear Regression is able to produce competitive models with a **Mean Squared Error (MSE)** and a **Mean Absolute Error (MAE)** on test data well below the +/−2°C threshold imposed by domain experts for the considered application scenarios. Furthermore, linear models are strongly preferred by domain experts because they are explainable and can be directly deployed on low-cost microcontrollers with ease.

3.3 Modelling Task as Multi-Objective Optimisation Problems

It is important to highlight that while the results above show that LR is a suitable technique for our data-driven dynamic thermal modelling scenario, collecting the 20 datasets (i.e., temperature profiles based on likely operational scenarios) was a very time-consuming exercise that also required specialised expertise. As such, there is a primary modelling imperative to discover if (and under which conditions) a more limited data collection stage can yield equally good LR models as this would significantly reduce modelling costs (especially when aiming to analyze more motor designs). Given an expected positive correlation between data availability and model accuracy, we opted to explore the aforementioned data collection inquiry through a set of three **multi-objective optimisation problems (MOOPs)**, each designed to provide a holistic answer to a modelling question grounded on the efficient usage of the 20 datasets ($DS_{01} \ldots DS_{20}$) in a manner that is likely to generate explainable thermal models:

Q1: *Which combination of datasets should be used to train an LR thermal model that is able to accurately estimate a given target temperature across all operational scenarios?* – Given the cost and complexity of collecting data, it would be important to know which load points are likely to help characterise the thermal behaviour of a particular motor component and the accuracy tradeoffs related to their usage during modelling.

Q2: *What EMA weights used for creating synthetic (input) features can improve the accuracy of LR thermal models for each of the six target temperatures in the context of reduced training sample availability?* – Instead of limiting synthetic feature generation to the three weights that capture short-, medium- and long-term trends as described in Section 3.1, the idea is to attempt to improve LR accuracy by extending the **Q1** modelling problem to include the identification of the best weights or combination of weights from a predefined range. Therefore, we generated 10 additional EMA weights by using the formula $\alpha_i = 0.001 \cdot 2^i,\ i \in \{0, 1, 2, \ldots, 9\}$ to capture a wider range of trends in the data. We then used the 10 weights to create 60 synthetic EMA features, one for each of the 6 features based on speed (v) and current (I). After replacing the 18 original EMA synthetic features with the 60 new ones, each of the 20 datasets we used for answering this question had a total of 66 input features. In terms of XAI characteristics, the optimisation of EMA weights can be seen as an attempt to mitigate the loss of LR (design) *transparency* induced by the initial arbitrary fixing of EMA settings.

Q3: *Which combination of datasets and EMA weights should be used when wishing to train accurate LR thermal models for all six target temperatures?* – Besides discovering the modelling trade-offs for a particular target temperature (i.e., answering **Q2**), it would also be very useful to investigate how optimal combination of datasets and EMA weights can be used to best model all six target temperatures via LR.

Formally, all three data modelling MOOPs that we aim to solve can be defined as:

$$Minimise\ F(x) = (f_1(x), f_2(x)) \tag{2}$$

where x is a n-dimensional vector of real-valued variables—i.e., $x_i \in D^n \subset \mathbb{R}^n, \forall 1 \leq i \leq n$; and $f_1 \in \mathbb{R}$ and $f_2 \in \mathbb{R}$ represent individual objectives:

- $f_1(x)$ = the total number of data samples in the training set encoded by x that are used for creating the LR model;
- $f_2(x)$ = the MAE or the MSE obtained by the trained LR model on the test set encoded by x.

As illustrated in Figure 4, in order to enable x to easily encode the training-test data split across our 20 datasets, we have formulated the three MOOPs as a typical 0,1 Knapsack problem, codified with real values [Russell and Norvig 2010].

In the case of MOOP1 – the problem designed to answer **Q1**, a candidate solution is a vector of 20 real-values between 0 and 1 (i.e., $x \in [0, 1)^{20}$) with the interpretation that each variable x_i represents its associated dataset DS_i. If $x_i \geq 0.5$, then DS_i is selected and added to the training set of the modelling experiment. On the other hand, if $x_i < 0.5$, DS_i is added to the test set of the modelling experiment. In order to evaluate $F(x)$, a counting of the total number of samples in the training set is performed (i.e., f_1) and an LR model is first trained on the training set and then tested on the test set to inform f_2. It is noteworthy that, since we are interested in the independent modelling of 6 different component temperatures, we are considering six instances of this problem: MOOP1$-T_w$, MOOP1$-T_{bi}$, MOOP1$-T_f$, and so on.

MOOP2 was formulated by adding 10 more variables to the decision vector used in MOOP1. Each new variable represents a predefined EMA weight. If a given weight is to be used (i.e., $x_i \geq 0.5, 21 \leq i \leq 30$), all the associated synthetic features (i.e., all 6 EMA features created with α_{i-21}) are used for training and testing the LR model that informs the accuracy of f_2. In other words, the usage of each EMA weight will add 6 independent variables to the resulting LR model.

MOOP3 is a variant of MOOP2 that features a minimax optimisation approach. For each candidate solution x we trained and tested independent LR models for all six target temperatures,

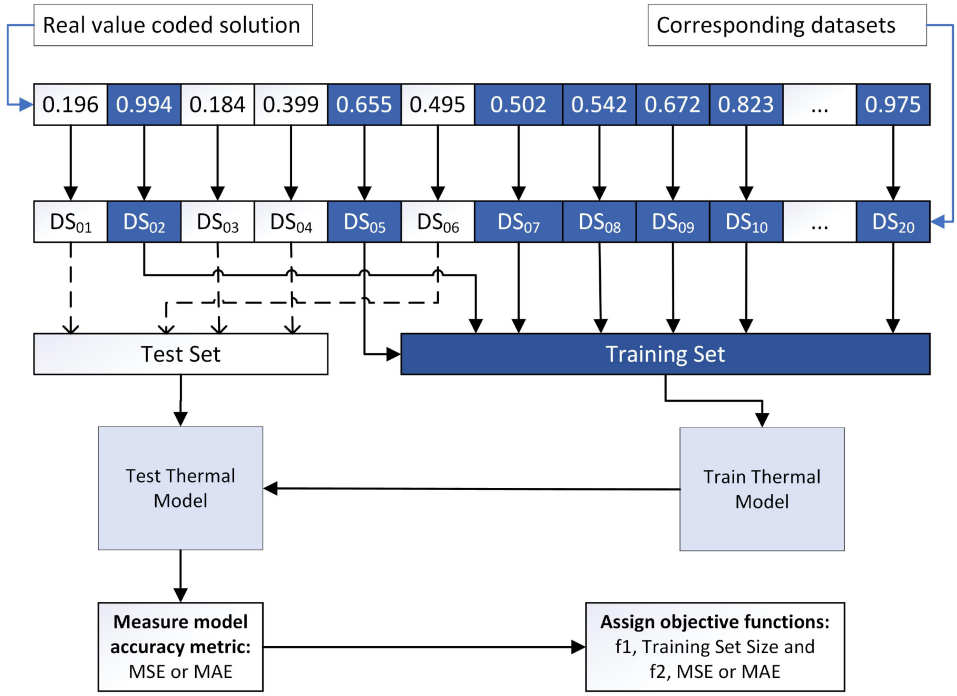

Fig. 4. Schematic of the proposed thermal modelling approach.

recording model test errors individually. We then defined $f_2(x)$ as the maximum test error observed across the six LR models, meaning that the accuracy objective of this MOOP aims to minimise the largest error across all component temperatures of interest.

3.4 Experimental Setup

Given the characteristics of our MOOPs (i.e., two objectives, unknown PF_t, medium number of decision variables), we opted to use the NSGA-II [Deb et al. 2002] solver, the second version of the **Nondominated Sorting Genetic Algorithm (NSGA-II)** is one of most widely used **multi-objective evolutionary algorithms (MOEAs)** and is known to be robust across different types of real-life and benchmark MOOPs. This means that NSGA-II is generally able to discover Pareto-optimal (PN) sets that very accurately approximate the true Pareto Front (PF_t) of the problem—i.e, the objective-space projection of all the optimal tradeoff solutions of the MOOP.

We applied NSGA-II with its standard genetic operators—i.e., **Simulated Binary Crossover (SBX)** [Deb et al. 1995] and polynomial mutation [Deb et al. 1996]—and we used the literature recommended settings for these operators: crossover probability rate of 0.8, crossover distribution index of 20, mutation probability of $1/n$ and a mutation distribution index of 20. Across all optimisation runs, we set both the population and offspring size to 200 and used a computational budget of 50,000 fitness evaluations, thereby evolving 250 generations. Given the stochastic nature of MOEAs, we initially carried out five independent repeats of each optimisation run. The limited number of runs is motivated by the fact that, even after parallelising the fitness evaluations, a typical optimisation would take 10-15 hours on a high-end PC. In the case of MOOP2, each modelling experiment was repeated 30 times in order to enable statistical significance testing of the importance of optimising the EMA weights.

Our numerical experiments integrated algorithm implementations from jMetalPy—a Python-based framework for multi-objective optimization with metaheuristics [Benítez-Hidalgo et al. 2019]—and Scikit-learn—a library for machine learning in Python [Pedregosa et al. 2011].

4 RESULTS AND INTERPRETATION

4.1 MOOP1: Optimising Data Requirements for Thermal Modelling

Figure 5 shows typical optimisation results for MOOP1. The top subplots show the training set size vs accuracy tradeoff for LR models of T_w when using the MAE (left) and the MSE (right) on the test set as model quality indicators. Similarly, the bottom subplots from Figure 5 indicate the sought modelling tradeoffs for T_{bi} —the other high priority component temperature. Across all subplots, we marked with black squares the Pareto-optimal solutions identified by NSGA-II (i.e., the objective space projection of the PN obtained at the end of the run). The x-axis is trimmed at 2 in light of the +/−2°C modelling accuracy constraint imposed by our thermal modelling scenario. Across both high-priority temperatures, test errors decrease with increasing training set size. However this decrease is very gradual and somewhat limited as models trained with fewer than 50,000 samples have MAE values smaller than 1°C and MSE values smaller than 1.5°C, while models trained with more than 200,000 samples have MAE and MSE values smaller than 0.5°C. On the one hand, this behaviour is expected because when there is a very limited set of samples to learn from, the LR model lacks the ability to properly model all the underlying patterns when presented with unseen temperature profiles. On the other hand, the fact that even models trained on less than 10% of the available data satisfy the accuracy constraint (i.e., generalise well) validates that LR is effective for modelling the dynamic thermal behaviour of the studied electrical machine. Thus, while not directly linked to explainability, the holistic view provided by the MOOP1 formulation and its associated results from Figure 5 reinforce user trust in the choice of regression model. We mention that these experiments were conducted for the four medium and low priority target temperatures as well and the results follow a very similar pattern.

Generally, MOOP1 modelling results show that an LR model trained on a subset of the original 20 datasets can be used to accurately predict target temperatures across different operational scenarios. For example, the Pareto optimal solution pointed with an arrow on the bottom left subplot from Figure 5 represents an LR model trained only using datasets DS_{03} and DS_{14} (i.e., $\approx$ 9.5% of all available data) that yielded a test MAE of 0.8610 on the other 18 datasets. This particular LR model is given in Equation (3) and, in light of its simplicity and accuracy, is a very interesting contender for installation on a microcontroller to estimate T_{bi} (the temperature of the inner ball bearing) when only provided with data regarding v (the rotor speed) and I (the electric current).

$$
\begin{aligned}
T_{bi} = & -(2.8913 \cdot v) - (0.2991 \cdot I) + (0.1384 \cdot I \cdot v) + (4.3532 \cdot v^2) - (2.2648 \cdot v^3) \\
& - (0.2466 \cdot I^2) - (4.2058 \cdot EMA_{0.001}(v)) - (7.1872 \cdot EMA_{0.005}(v)) \\
& + (5.6999 \cdot EMA_{0.04}(v)) + (1.7672 \cdot EMA_{0.001}(I)) + (4.1274 \cdot EMA_{0.005}(I)) \\
& - (2.1350 \cdot EMA_{0.04}(I)) + (1.3143 \cdot EMA_{0.001}(I \cdot v)) - (7.0290 \cdot EMA_{0.005}(I \cdot v)) \\
& + (2.3738 \cdot EMA_{0.04}(I \cdot v)) + (9.5421 \cdot EMA_{0.001}(v^2)) + (24.8746 \cdot EMA_{0.005}(v^2)) \\
& - (12.6139 \cdot EMA_{0.04}(v^2)) - (5.0606 \cdot EMA_{0.001}(v^3)) - (14.2362 \cdot EMA_{0.005}(v^3)) \\
& + (7.7438 \cdot EMA_{0.04}(v^3)) + (0.6009 \cdot EMA_{0.001}(I^2)) + (3.0972 \cdot EMA_{0.005}I^2)) \\
& + (1.3959 \cdot EMA_{0.04}(I^2)) + 42.3631
\end{aligned}
\tag{3}
$$

We proceeded to compare the performance of LR models for T_w and T_{bi} trained only using DS_{03} and DS_{14} with the non-linear alternatives considered in Section 3.2. To make this comparison, we first applied the previously outlined strategies for identifying the best parameters for each non-linear modelling technique when considering only the 22,862 samples from the two training datasets. We then trained the non-linear models using all the 22,862 samples and finally tested

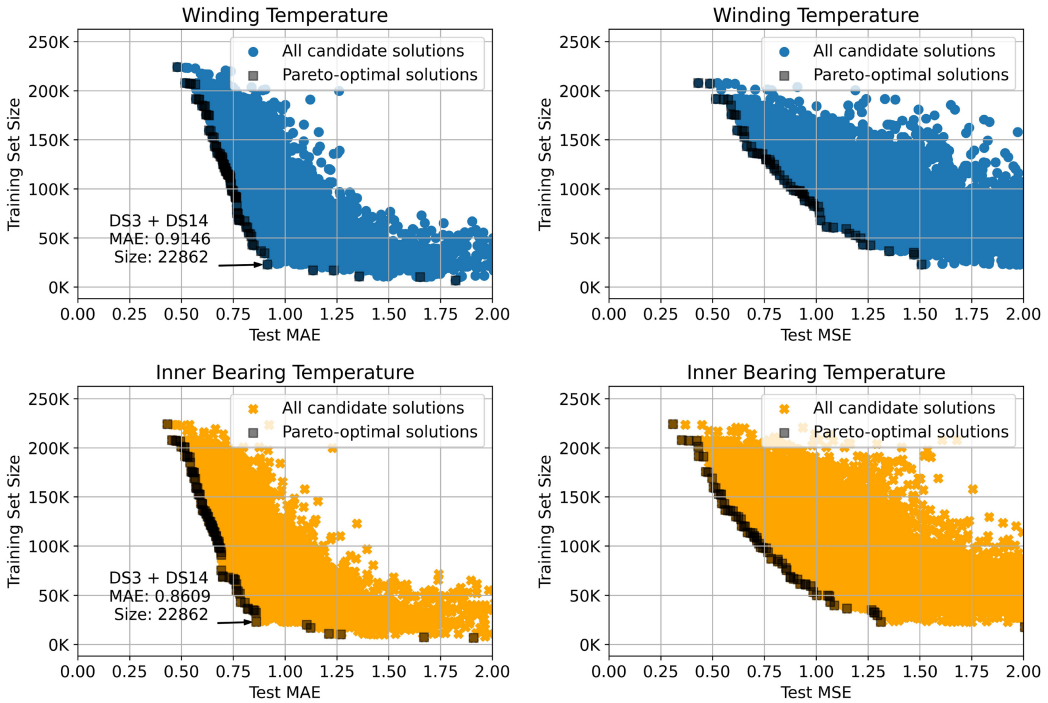

Fig. 5. Pareto fronts (black squares) of single NSGA-II optimisation runs on MOOP1 when aiming to model T_w (top subplots) and T_{bi} (bottom subplots).

Table 2. Comparative Performance on Test Data when Only Training using DS_{03} and DS_{14}

Qual. indicator	RF		KNN		ANN		LR	
	T_w	T_{bi}	T_w	T_{bi}	T_w	T_{bi}	T_w	T_{bi}
MSE	4.186	4.520	15.53	6.650	5.661	1.751	**1.508**	**1.312**
MAE	1.362	1.315	2.392	1.689	1.492	0.972	**0.915**	**0.861**
R^2	0.965	0.858	0.870	0.788	0.952	**0.945**	**0.987**	0.958

The best result for each (component temperature, quality indicator) pair is highlighted in bold font.

them on the the remaining 217,338 samples from the other 18 datasets. The results are shown in Table 2 and indicate that, when compared with the preliminary results from Table 1, the MAE and MSE performance degradation is an order of magnitude higher for the non-linear approaches. This can be interpreted as further evidence towards the robustness and overall suitability of LR models for our considered modelling tasks, especially when aiming to reduce data collection requirements.

4.2 MOOP2: Optimising the EMA Weights Used for Synthetic Feature Generation

In MOOP2, we included EMA weights into the optimisation and the obtained results follow a similar pattern as those obtained for MOOP1. In the two subplots from Figure 6, we illustrate all the Pareto-optimal solutions discovered by NSGA-II for MOOP1 and MOOP2 across the five initial independent runs when modelling T_w. Graphically, it is clear that test errors decrease when both training set composition and EMA weights are optimised and as a result the Pareto fronts associated with MOOP2 are shifted to the left. In order to further investigate this empirical observation,

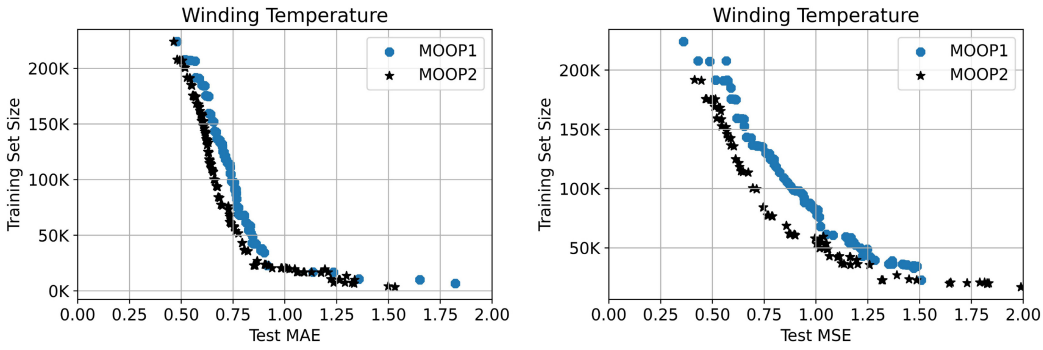

Fig. 6. All end-of-the-run Pareto-optimal solutions for MOOP1 (datasets optimisation) and MOOP2 (datasets + EMA weights optimisation) across five independent runs on each problem that aimed to model the winding temperature (T_w).

we carried out 25 more independent optimisation runs and proceeded to quantitatively measure the quality of the obtained Pareto fronts. Several specialised indicators are commonly used for this task: the generational distance [Van Veldhuizen and Lamont 1998], the inverse generational distance [Coello et al. 2007], the epsilon indicator [Zitzler et al. 2003] and the hypervolume indicator [Zitzler and Thiele 1998].

We chose to use the hypervolume indicator (Hv) as our unary PF quality measure because it is widely accepted in the MOEA community, has a theoretical proof of a monotonic convergence behaviour, and can be easily used on problems with an unknown PF_t. This is because $Hv(PF_c)$ measures the size of the objective space that PF_c dominates when considering an anti-optimal reference point [Zitzler and Thiele 1998]. Based on this, larger Hv values are preferred, but in order to make the numerical values more meaningful, computing the relative hypervolume as $Hr(PF_c) = \frac{Hv(PF_c)}{Hv(PF_t)}$ is advisable. In our case, as PF_t is unknown, we have decided to assume it only contains the ideal point (0,0) that would denote an LR model that requires 0 training data and yields 0 errors. Conversely, the anti-optimal reference point was set at (5, 240200), denoting a hypothetical LR model that is trained using 100% of the data but falls well out of acceptable accuracy thresholds.

Across 30 independent runs aimed at modelling T_w, we obtained:

- an average Hr of 78.00% and a median Hr of 77.96% in the case of MOOP1 (i.e., when only optimising the temperature profiles used for training);
- an average Hr of 81.59% and a median Hr of 81.62% in the case of MOOP2 (i.e., when optimising both profiles and EMA-based synthetic features).

This general improvement of modelling outcomes suggested by the difference in Hr central tendency indicators between MOOP1 and MOOP2 was confirmed as statistically significant by a one-sided Mann-Whitney U test [Mann and Whitney 1947] with a 0.01 significance level (p-value = $1.5099 \cdot 10^{-11}$). This means that we can say with 99% confidence that the inclusion of EMA weights in the optimisation improves the data requirement vs. accuracy tradeoffs of our LR thermal models for T_w. The impact of this decision on model *explainability* is discussed at length in Section 4.4.

4.3 MOOP3: Simultaneously Optimising All Six Target Temperatures

Figure 7 shows a typical optimisation result for MOOP3 where, given our minmax approach described at the end of Section 3.3, for each evaluated solution, the color and shape (as per the legend)

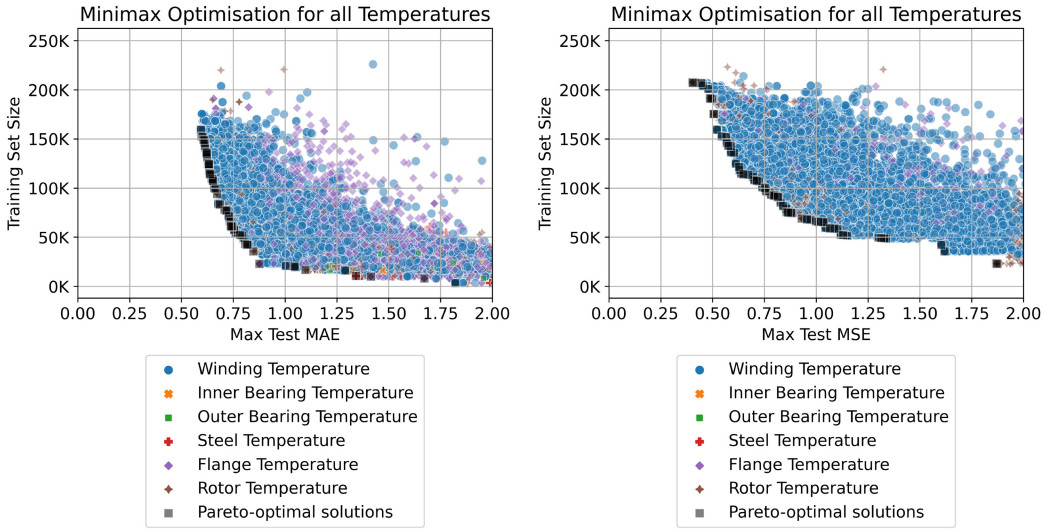

Fig. 7. Typical MOOP3 optimisation result.

correspond to the target temperature for which the solution's maximum LR test error was obtained. The winding temperature (T_w) error is dominant across all evaluated solutions and among plotted Pareto optimal solutions, the max error value is associated with (i) T_w in 49/62 of cases, T_r in 10/62 of cases, and T_f in 3/62 of cases (for MAE) and with (ii) T_w in 53/59 of cases, T_r in 5/59 of cases, and T_f in 1/59 of cases (for MSE).

Based on this, we can infer that a combination of datasets (i.e., sample temperature profiles) and EMA weights that can lead to an accurate LR model for predicting the winding temperature will equally yield accurate LR models for predicting all six component temperatures under a wide range of operational scenarios. This observation and the high-priority modelling status motivates the T_w significance testing focus in Section 4.2.

4.4 Balancing Model Accuracy and Explainability

We are aware that an excessive use of EMA synthetic features (in solution to MOOP2 and MOOP3) will increase the complexity of the LR models thus compromising our stated objective of obtaining simple and explainable models that can help electrical engineers gain insights related to the dynamic thermal behavior of the studied electrical machine. For example, in Figure 8, we re-plot all the T_w-based Pareto optimal solutions from the MOOP3 run depicted in Figure 7 with a marker size proportional to the size of each LR model of T_w. These results indicate that the improved accuracy brought by including EMA-weights in the multi-objective optimisation tends to come at the expense of generating larger (i.e., more complex) models when increasing the amount of training data. This is especially obvious when using MSE as an optimisation goal and is likely due to the fact that the usage of the same quadratic loss function within the MSE and LR formulae enables a larger set of EMA-weights to bring marginal modelling improvements when training on larger sets of temperature profiles. When the loss functions used in the optimisation and model training are well aligned but not identical (i.e., when $f_2(x)$ is based on MAE), the increase of optimal model size is more subdued.

The fact that complexity increase affects MAE and MSE modelling differently is also evidenced by the plots in Figure 9 that display the comparative performance of the Pareto optimal LR models

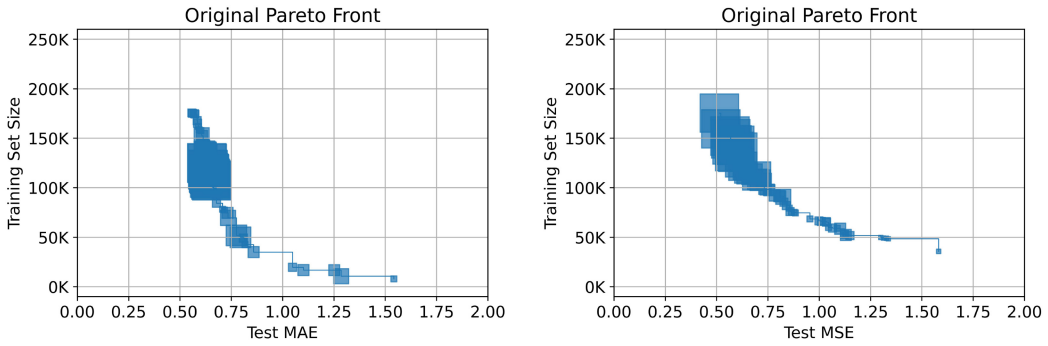

Fig. 8. Size vs Pareto Front (PF) position of T_w-based solutions at the end of a MOOP3 run.

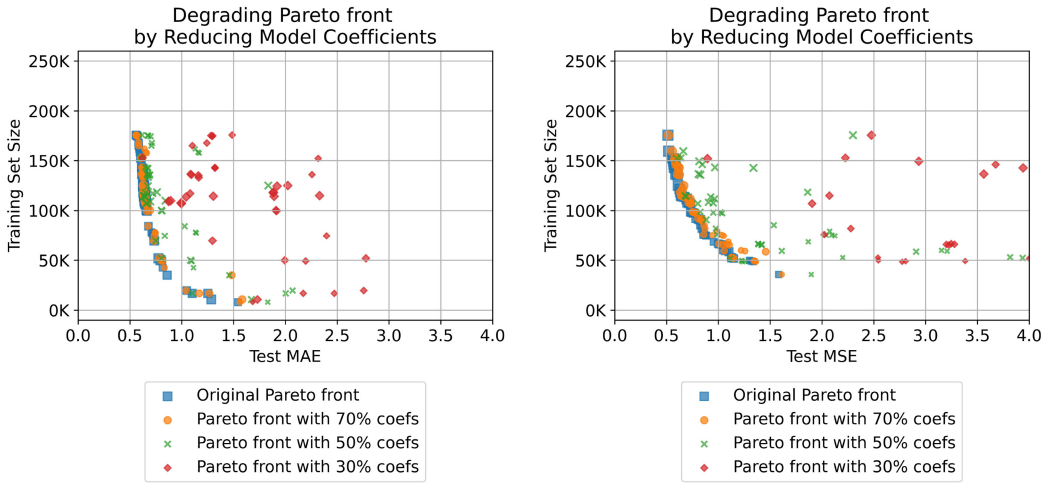

Fig. 9. Degraded Pareto fronts.

from Figure 8 after a step-wise regularisation procedure that removes 30%, 50%, and 70% of the original regression model coefficients in decreasing order of their importance (i.e., absolute value).

Regularisation results indicate that a reduction of LR model size (complexity) by 50% to 70% affects MSE optimal models more (i.e., they determine larger error increases). The fact that a 30% reduction of model sizes appears to have a negligible effect on estimated accuracy for most optimal models can be explained by our MOOP formulation described in Section 3: when an EMA weighting is selected, six new synthetic features (corresponding to two original + four expert-suggested base features) are created and all six features will feature in the final LR thermal model even if just one feature has a meaningful contribution to improving model accuracy. This approach was a design tradeoff itself as:

- we wished to limit the size of our MOOPs. By allowing the multi-objective solver to select individual EMA synthetic features, the sizes of MOOP2 and MOOP3 search space would increase to 70 instead of 20 – likely requiring a more complicated solver + parameterisation selection process alongside extended run-times;
- we wanted to aim the modelling exercise towards identifying EMA weights that capture temporal trends that are relevant for more multiple base features as these weights could provide more insights to electrical engineers (thus improving overall *explainability*).

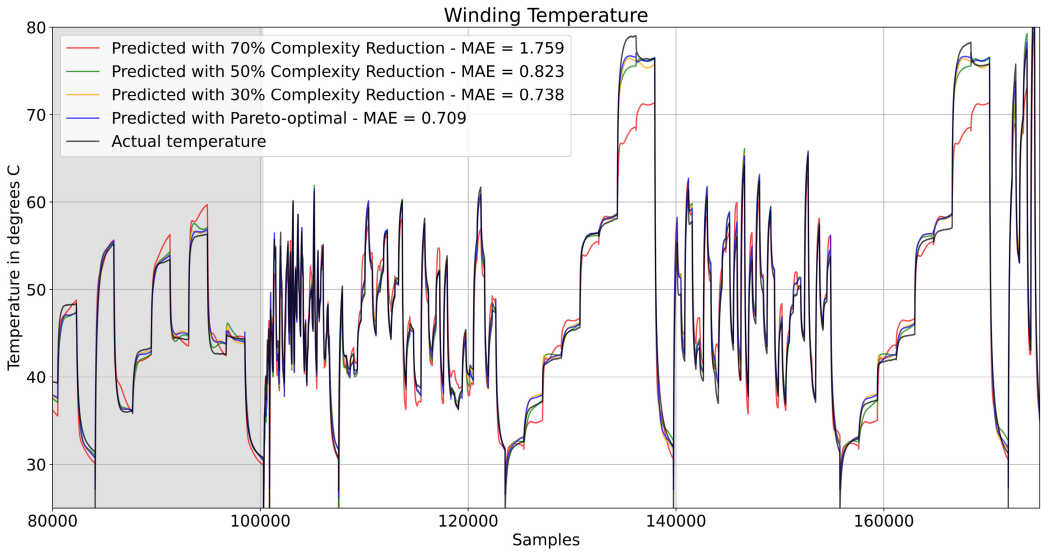

Fig. 10. Performance of a T_w model for MOOP3 at different levels of regularisation across several operational scenarios (the grey area denotes part of the samples used during model training).

Meaningful EMA weights can be identified by domain experts that analyse relative temperature profiling differences on LR models where a reduction of complexity is more strongly correlated to a corresponding reduction of global and/or local modelling accuracy—e.g., the T_w model from Figure 10.

Regarding the relative modelling performance shown in Figure 10, it is noteworthy that features that are in the 50% to 70% range of importance (based on their associated absolute coefficient values in the original Pareto optimal LR model) seem crucial for correctly modelling temperature peaks associated with constant medium and high utilisation scenarios. Conversely, the least important 50% of original model features have an incremental, but overall very limited, impact on general modelling performance. These observations indicate that by further tailoring the regularisation procedure (e.g., making it more fine grained or dependent on the relative loss of global/local accuracy across the 20 analysed scenarios), the explainability of the original model could be enhanced by constructing a more detailed mapping of features or groups of features to particular thermal profiling capabilities. This in turn would give decision makers a clear view of all the modelling tradeoffs associated with a given Pareto optimal thermal model: training costs vs accuracy vs explainability.

5 CONCLUSIONS AND FUTURE WORK

The present research demonstrates how three 0,1 Knapsack multi-objective formulations of data modelling tasks coupled with the usage of an effective evolutionary solver (i.e., NSGA-II) can be used to outline optimal costs vs accuracy tradeoffs when aiming to discover high-quality **Linear Regression (LR)** models that can estimate the dynamic thermal behaviour of six electrical motor components under various operational scenarios. Case study results indicate that the ability to generate highly explainable models coupled with the holistic data modelling perspective provided by our multi-objective approach provides electrical engineers with useful data-driven insights regarding the thermal profile of the studied electrical machine.

In particular, we have shown that by creating synthetic features using **Exponential Moving Averages (EMAs)** with optimised weights, one can obtain highly accurate LR models, even when drastically reducing the required amount of training data, but this does impact explainability by increasing the complexity (i.e., size) of high-performing LR models. To alleviate this issue, we demonstrate how a very basic step-wise regularisation technique can be applied to reduce complexity (with minimal impact on accuracy) and improve explainability by facilitating a domain relevant mapping of features to modelling capabilities.

Further work will aim to build on present results by testing different methods of constraining and/or reducing linear model complexity. We will primarily focus on well-known regularisation techniques (i.e., ridge, lasso, elastic net) and on effective ways of directly integrating model complexity as an optimisation objective in its own right. We envision that an extension of the proposed multi-objective data-driven modelling approach to other ML paradigms known to display an accuracy vs explainability tradeoff (e.g., symbolic regression, decision trees) will define a secondary future work stream. Finally, it would be of particular interest to compare our results with those obtained by multi-objective **Genetic Programming (GP)** approaches [Burlacu et al. 2019; Kommenda et al. 2016] given the ability of the latter to also explore tradeoffs between evolving simpler (i.e., more interpretable) or more numerically accurate symbolic regression models.

REFERENCES

Antonio Benítez-Hidalgo, Antonio J. Nebro, José García-Nieto, Izaskun Oregi, and Javier Del Ser. 2019. jMetalPy: A Python framework for multi-objective optimization with metaheuristics. *Swarm and Evolutionary Computation* 51 (2019), 100598. https://doi.org/10.1016/j.swevo.2019.100598

Aldo Boglietti, Andrea Cavagnino, David Staton, Martin Shanel, Marjus Mueller, and Carlos Mejuto. 2009. Evolution and modern approaches for thermal analysis of electrical machines. *IEEE Transactions on Industrial Electronics* 56 (2009), 871–882. Issue 3. https://doi.org/10.1109/TIE.2008.2011622

Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/a:1010933404324

Bogdan Burlacu, Gabriel Kronberger, Michael Kommenda, and Michael Affenzeller. 2019. Parsimony measures in multi-objective genetic programming for symbolic regression. (2019).

B. J. Chalmers and E. Spooner. 1999. An axial-flux permanent-magnet generator for a gearless wind energy system. *IEEE Transactions on Energy Conversion* 14, 2 (June 1999), 251–257. https://doi.org/10.1109/60.766991

Anurag Choudhary, Deepam Goyal, Sudha Letha Shimi, and Aparna Akula. 2018. Condition monitoring and fault diagnosis of induction motors: A review. *Archives of Computational Methods in Engineering* 26, 4 (Sep 2018), 1221–1238. https://doi.org/10.1007/s11831-018-9286-z

Carlos A. Coello Coello, Gary B. Lamont, and David A Van Veldhuizen. 2007. *Evolutionary Algorithms for Solving Multi-objective Problems*. Springer.

T. Cover and P. Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13, 1 (Jan 1967), 21–27. https://doi.org/10.1109/tit.1967.1053964

Kalyanmoy Deb and Ram Bhushan Agrawal. 1995. Simulated binary crossover for continuous search space. *Complex Systems* 9, 2 (1995), 115–148.

Kalyanmoy Deb and Mayank Goyal. 1996. A combined genetic adaptive search (GeneAS) for engineering design. *Computer Science and Informatics* 26 (1996), 30–45.

Kalyanmoy Deb, Amrit Pratap, Sameer Agarwal, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation* 6 (2002), 182–197. Issue 2. https://doi.org/10.1109/4235.996017

Simon Haykin. 1999. *Neural Networks: A Complrehensive Foundation* (2nd ed.). Pearson Prentice Hall.

Charles C. Holt. 2004. Forecasting seasonals and trends by exponentially weighted moving averages. *International Journal of Forecasting* 20, 1 (2004), 5–10.

Wilhelm Kirchgässner, Oliver Wallscheid, and Joachim Böcker. 2019. Deep residual convolutional and recurrent neural networks for temperature estimation in permanent magnet synchronous motors. In *Proceedings of the 2019 IEEE International Electric Machines & Drives Conference (IEMDC)*. 1439–1446. https://doi.org/10.1109/IEMDC.2019.8785109

Michael Kommenda, Gabriel Kronberger, Michael Affenzeller, Stephan M. Winkler, and Bogdan Burlacu. 2016. Evolving simple symbolic regression models by multi-objective genetic programming. *Genetic Programming Theory and Practice XIII* (2016), 1.

Michael H. Kutner, Christopher J. Nachtscheim, John Neter, and William Li. 2005. *Applied Linear Statistical Models* (5th ed.). McGraw-Hill Irwin.

Zachary C. Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

H. B. Mann and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* 18, 1 (Mar 1947), 50–60. https://doi.org/10.1214/aoms/1177730491

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing* 73 (2018), 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

J. N. Nash. 1997. Direct torque control, induction motor vector control without an encoder. *IEEE Transactions on Industry Applications* 33, 2 (Mar 1997), 333–341. https://doi.org/10.1109/28.567792

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 1135–1144. https://doi.org/10.1145/2939672.2939778

Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. 2020. Explainable machine learning for scientific insights and discoveries. *IEEE Access* 8 (2020), 42200–42216. https://doi.org/10.1109/access.2020.2976199

Stuart Russell and Peter Norvig. 2010. *Artificial Intelligence: A Modern Approach* (3rd ed.). Pearson Education, Inc.

David A. Van Veldhuizen and Gary B. Lamont. 1998. *Multiobjective Evolutionary Algorithm Research: A History and Analysis*. Technical Report TR-98-03. Department of Electrical and Computer Engineering, Graduate School of Engineering, Air Force Institute of Technology, Wright-Patterson AFB, Ohio.

Daniel Wöckinger, Gerd Bramerdorfer, Stephan Drexler, Silvio Vaschetto, Andrea Cavagnino, Alberto Tenconi, Wolfgang Amrhein, and Frank Jeske. 2020. Measurement-based optimization of thermal networks for temperature monitoring of outer rotor PM machines. In *Proceedings of the 2020 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 4261–4268. https://doi.org/10.1109/ECCE44975.2020.9236388

Daniel Wöckinger, Gerd Bramerdorfer, Silvio Vaschetto, Andrea Cavagnino, Alberto Tenconi, Wolfgang Amrhein, and Frank Jeskey. 2021. Approaches for improving lumped parameter thermal networks for outer rotor SPM machines. In *Proceedings of the 2021 IEEE Energy Conversion Congress and Exposition (ECCE)*. IEEE, 3821–3828. https://doi.org/10.1109/ECCE47101.2021.9594930

Eckart Zitzler and Lothar Thiele. 1998. Multiobjective optimization using evolutionary algorithms — A comparative case study. In *Lecture Notes in Computer Science*. Springer Berlin, 292–301. https://doi.org/10.1007/bfb0056872

Eckart Zitzler, Lothar Thiele, Marco Laumanns, Carlos M. Fonseca, and Viviane Grunert da Fonseca. 2003. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation* 7, 2 (Apr 2003), 117–132. https://doi.org/10.1109/tevc.2003.810758

Alexandru-Ciprianan Zăvoianu, Martin Kitzberger, Gerd Bramerdorfer, and Susanne Saminger-Platz. 2020. On modeling the dynamic thermal behavior of electrical machines using genetic programming and artificial neural networks, In *Computer Aided Systems Theory – EUROCAST 2019* (Cham), R. Moreno-Díaz, F. Pichler, and A. Quesada-Arencibia (Eds.). Springer International Publishing, 319–326.

Received 15 November 2022; revised 16 March 2023; accepted 27 April 2023