Oral Temperature Prediction with Infrared Thermography Using Tree-based Regression Models

Zhixin Mao, Ziyi Ding

Introduction

Body temperature is a very common measurement for many diseases such as Severe Acute Respiratory Syndrome (SARS) in 2003 [1] and Coronavirus in 2019 (COVID-19) [2]. There are many important body sites to measure body temperature, but some of them are not convenient to access. Infrared thermography (IRT) is a technology that detects infrared energy emitted from an object, converts it to temperature, and displays an image of temperature distribution [3]. This is one of the primary devices commonly used in practice for measuring body temperature and testing disease. The utility of infrared thermography lies in its non-contact nature of temperature measurement, which represents an essential tool in the medical field. However, the evaluation of IRT clinical accuracy is not fully tested due to its relevance to the variability of the real world [4]. According to Ismael's study, many factors can impact the measurement, such as gender, skin humidity, or hair density [5]. It is not practical to control for all factors, but analyzing their relationships with IRT is very important for enhancing our understanding and application of this technology in a clinical setting. Our primary intent of this study was to apply different tree-based approaches for predicting oral temperature and finding the relationship between IRT and factors.

In the analysis of datasets where outcomes are continuous, and predictors exhibit high collinearity, tree-based regression models, such as decision trees, random forests (RF), and gradient-boosting trees, offer significant advantages over traditional linear regression models. Tree-based models can capture complex, non-linear relationships without assuming a specific form of interaction between predictors and the outcome [6]. Unlike linear regression, which requires explicit specification of interaction terms and is sensitive to collinearity, tree-based models naturally incorporate interactions and are robust to collinearity among predictors [7]. In addition, by automatically capturing interaction effects within their hierarchical structure, these models eliminate the need for prior assumptions to identify potential interactions among predictors. These advantages make tree-based regression models particularly well-suited for analyzing datasets where the underlying relationships between variables are complex and the presence of collinearity among predictors poses significant analytical challenges.

In this paper, considering complex predictor-outcome relationships and high collinearity among predictors, we applied four tree-based regression models, including bagging, RF, gradient-boosting trees, and Bayesian additive regression trees (BART) to compare selected predictors via different regression trees. In addition, we used selected predictors to predict oral temperature in both fast mode and monitor mode and measured the predictive performance among these four regression models.

State of the Art

In the previous study of Wang et al.[4], a large-scale clinical analysis of fever-screening thermography was done by the Health Center of the University of Maryland (UMD) to collect clinical information from more than 1000 subjects measured within a wide room temperature range of 20-29 degrees Celsius. Their goal was to compare different methods for IRT calibration, find the best practices for evaluating the clinical performance of IRT, and compare the results of IRT with a non-contact infrared thermometer (NCIT). In their study, they found a high correlation and non-linear relationship between their variables, but they only considered three regression methods, which are weighted linear regression, binning method, and piecewise regression. Their analysis indicated that there was no clear optimal method that could improve all clinical accuracy metrics. Given the limitation of linear regression approaches, testing more flexible tree-based methods on the prediction of oral temperature by using this same dataset is necessary.

Another similar work to predict non-contact core body temperature by testing regression models has been done by Chayabhan et al.[8]. They proposed a trained model prediction using IR-measured facial feature temperatures to predict reference body temperature. To improve body temperature measurement accuracy, they investigated five different types of regression models: linear regression models, regression trees, support vector machines, ensembles of trees, and Gaussian Process regression. Even though their study has concluded that the linear regression model showed the lowest minimum-root-mean-square error compared with reference temperatures, tree-based models also had good performances on temple plane and eye regions of interest.

Materials and methods

1 Dataset

1.1 Data Souce

The raw data was collected by a clinical study at the Health Center of the University of Maryland (UMD) at College Park from November 2016 to May 2018 [9]. There were two IRTs used in this clinical study– one was FLIR (IRT-1) and the other was ICI (IRT-2). In this analysis, we only focused on FLIR measurement in round 1. There were two separate groups for the infrared images–one was Group 1 with an ambient ranging from 20.0 to 24.0 ℃ and the other was Group 2 with an ambient ranging from 24.0 to 29.0 ℃. We analyzed these two group data together in this study. This dataset contained 1020 subjects and 38 variables which contains a total of 26 facial temperature variables and demographic information including gender, age, ethnicity, ambient temperature, humidity, and measuring distance. The targets are average oral temperatures under two operation modes (fast mode and monitor mode) of the oral thermometer. For modeling, we did not include subject, cosmetic, time, and date in our analysis. (The details for more data information are attached in supplementary).

1.2 Data Process

In this dataset, not every subject has a full record of all measurements, so we dropped those subjects containing missing values. In total, there are 1001 subjects left for further analysis. For consistency, we combined the two age groups from 21-25 and 26-30 to one age group from 21-30. To evaluate our modeling, the subjects were randomly separated into training and testing sets. The training set (80% of subjects) contains 600 subjects and the testing set (20% of subjects) contains 401 subjects.

2 Tree-based Regression Methods

In our methodology, we employed tree-based regression models due to their versatility and robustness in handling complex datasets. Decision trees are the foundation for several ensemble techniques that enhance prediction accuracy and model stability. Bagging reduces variance by averaging multiple decision trees trained on different samples of the data. RF extends bagging by also sampling features, further improving accuracy and robustness to overfitting. Boosting sequentially fits trees on modified versions of the data, increasing predictive performance. Lastly, BART combines the principles of Bayesian statistics with decision trees, offering a probabilistic approach to modeling and inference.

We built regression models to predict both oral temperature measured in fast mode (aveOralF) and oral temperature measured in monitor mode (aveOralM) and measured their performance utilizing the test set.

2.1 Bagging

Bagging, also called bootstrap aggregating, is an ensemble learning method to improve accuracy and model stability [10]. This method can be applied to both regression and classification problems with the ability to avoid overfitting of data and deal with bias-variance trade-offs. Bagging is usually based on multiple decision trees and averaged with the trained model to get a more accurate estimate.

The usual procedure for how the bagging works begins by creating multiple bootstrap samples [11]. A bootstrap sample is a randomly selected subset of data chosen with replacement from the original dataset. This means individual observations can appear more than once in the same bootstrap sample or across different samples. For each bootstrap sample, a separate model is trained. Since the data in each sample varies, the resulting models will be different, capturing different patterns in the data. After training all models, bagging makes predictions by aggregating the outcomes of these individual models. For the regression problem, the final result is usually the mean of all outputs. For the classification problem, the final result is the majority vote.

2.2 Random Forest

Like bagging, RF is also an ensemble learning method widely used for both regression and classification tasks [12]. It builds upon the concept of bagging by using a collection of decision trees to create a more powerful model and combining the results of each model to make predictions. Unlike bagging, random forests use a modified tree-learning algorithm that selects a random subset of features.

The algorithm also starts with taking a random sample from training data with replacement. For each bootstrap sample, a decision tree grows. At each split in the tree, instead of considering all features to make the best split, a random subset of predictors is selected without replacement. After each tree is constructed, we drop the out-of-bag data down the tree and store the output. For regression, the prediction of the model is the average of all trees' outputs. For classification, the prediction is the class with the most votes.

Random forest effectively reduces variance and bias by combining bootstrap samples and random draws of predictors. Unlike bagging, the fitted values across trees in RF are more independent because it only considers a random subset of predictors at each split.

2.3 Gradient-boosting Trees

Boosting is another approach for improving the predictions resulting from a decision tree. Like bagging, boosting is a general approach that can be applied to many statistical learning methods for regression or classification trees. Unlike bagging, where trees are built independently, boosting involves a sequential process where the development of each tree is heavily influenced by those previously grown.

The boosting works via the following procedure [13]: firstly, create multiple copies of the original training data set using the bootstrap, and assign equal weights to all observations in the dataset. Secondly, grow each tree sequentially, using the errors of the prior tree to adjust observation weights. Mispredicted observations receive increased focus. Thirdly, after fitting each tree, update weights to emphasize errors, guiding the next tree's focus. Finally, aggregate the sequential trees into a final model using a weighted average based on accuracy, improving prediction through iterative correction.

Boosting differs from constructing a single large decision tree by adopting a gradual learning process. This method avoids the risk of overfitting associated with fitting a large decision tree too closely to the data [14]. Generally, statistical learning techniques that adopt a slower learning rate tend to achieve better performance.

2.4 Bayesian Additive Regression Trees

The BART model is composed of a sum-of-trees component for modeling and a regularization prior that governs the model parameters. The sum-of-trees model represents an additive framework incorporating multivariate elements, offering a more natural way to include interaction effects than generalized additive models [15]. BART employs a specialized Bayesian backfitting Markov Chain Monte Carlo (MCMC) [16] technique to fit the sum-of-trees model, iteratively building and adjusting for residuals in succession. While sharing conceptual similarities with gradient boosting, BART distinguishes itself by moderating the influence of individual trees through the application of a prior and by employing Bayesian backfitting for iterative fitting across a predetermined number of trees.

The BART works via the following procedure [16]: firstly, specify a model where multiple trees collectively capture data relationships. Secondly, use priors to limit individual tree influence, ensuring no single tree overly dominates the prediction and maintaining model balance. Thirdly, employ a Bayesian backfitting MCMC method for model fitting. This process fits trees sequentially, adjusting for the cumulative residuals left by all previously fitted trees. Finally, fit the model iteratively using a fixed number of trees, refining the model with each iteration by fitting each new tree to the current residuals. Continue the iterative fitting process until the model converges.

BART is essentially a Bayesian nonparametric strategy that applies a highly influential prior distribution to fit a model rich in parameters. Compared to models based on a single tree, the sum-of-trees approach more effectively captures additive effects.

3 Evaluation

For modeling, we applied 5-fold cross-validation to optimize parameters for those four tree-based models. The 5-fold cross-validation is very efficient for model selection and hyperparameter tuning with a limited sample size. For evaluation, we applied the testing data to each trained model and calculated root mean squared error (RMSE) as our metrics for comparing their performances.

Results

1 Descriptive Statistics

The demographic information for study subjects is summarized in Table 1. There are 601 females and 400 males in this study, and about 95% of subjects are under 30 years of age. Their ethnicities mostly come from Asian, Black/African American, and White.

Table 1. Overview of Categorical Variables

Variable		IRT	Variable		IRT
Gender	Female	601			
	Male	400			
Age	18-20	524	Ethnicity	American Indian	4
	21-30	423		Asian	255
	31-40	31		Black/African-American	143
	41-50	9		Hispanic/Latino	57
	51-60	11		Multiracial	49
	>60	3		White	493

Table 2. Overview of Continuous Variables

Variable	Min, Max	Mean (SD)	Variable	Min, Max	Mean (SD)
Distance	0.54, 79.00	0.73 (2.48)	T_LC_Max1	33.79, 38.21	35.60 (0.57)
Humidity	9.90, 61.20	28.76 (13.07)	T_LC_Wet1	33.12, 37.93	35.39 (0.62)
LCC1	32.57, 37.91	35.13 (0.65)	T_Max1	33.85, 38.52	35.82 (0.53)
Max1L13_1	33.55, 38.16	35.54 (0.59)	T_OR1	32.18, 37.75	35.29 (0.70)
Max1R13_1	33.35, 39.52	35.53 (0.61)	T_OR_Max1	32.18, 37.76	35.33 (0.70)
RCC1	32.78, 38.31	35.18 (0.65)	T_RC1	33.34, 38.51	35.59 (0.58)
T_FHBC1	30.26, 37.27	34.35 (0.77)	T_RC_Dry1	33.35, 38.51	35.51 (0.60)
T_FHCC1	30.02, 37.18	34.44 (0.77)	T_RC_Max1	33.35, 38.52	35.62 (0.58)
T_FHC_Max1	31.63, 37.58	35.00 (0.64)	T_RC_Wet1	33.30, 38.41	35.48 (0.61)
T_FHLC1	30.21, 37.58	34.45 (0.77)	T_atm	20.20, 29.10	24.21 (1.35)
T_FHRC1	29.54, 37.24	34.44 (0.79)	T_offset1	-0.76, 3.58	0.96 (0.42)
T_FHTC1	27.70, 37.44	34.43 (0.91)	aveAllL13_1	31.54, 37.78	34.91 (0.72)
T_FH_Max1	33.08, 38.02	35.36 (0.56)	aveAllR13_1	29.95, 37.74	34.79 (0.80)
T_LC1	33.78, 38.16	35.57 (0.57)	**aveOralF**	**35.75, 39.60**	**36.98 (0.39)**
T_LC_Dry1	33.67, 38.16	35.54 (0.58)	**aveOralM**	**35.54, 40.34**	**37.03 (0.50)**

Note: the bold variables represent the targets.

The summary of continuous variables is shown in Table 2 and the correlation matrix between continuous variables is plotted in Figure 1. Most notably, humidity in summary has a wide range and variance, which is explained by the fact that this study lasted for a long time covering all four seasons. In addition, there are very high correlations between variables. Especially, Max1R13_1 is highly correlated with T_RC1, T_RC_Dry1, T_RC_Wet1, T_RC_Max1, and RCC1, and Max1L13_1 is highly correlated with T_LC1, T_LC_Dry1, T_LC_Wet1, T_LC_Max1, canthiMax1 and canthi4Max1. Their correlations are all larger than 0.9. There are also high correlations between targets and predictors. For example, the correlation between aveOralM and T_Max1 is 0.79, and the correlation between aveOralF and T_Max1 is 0.69.

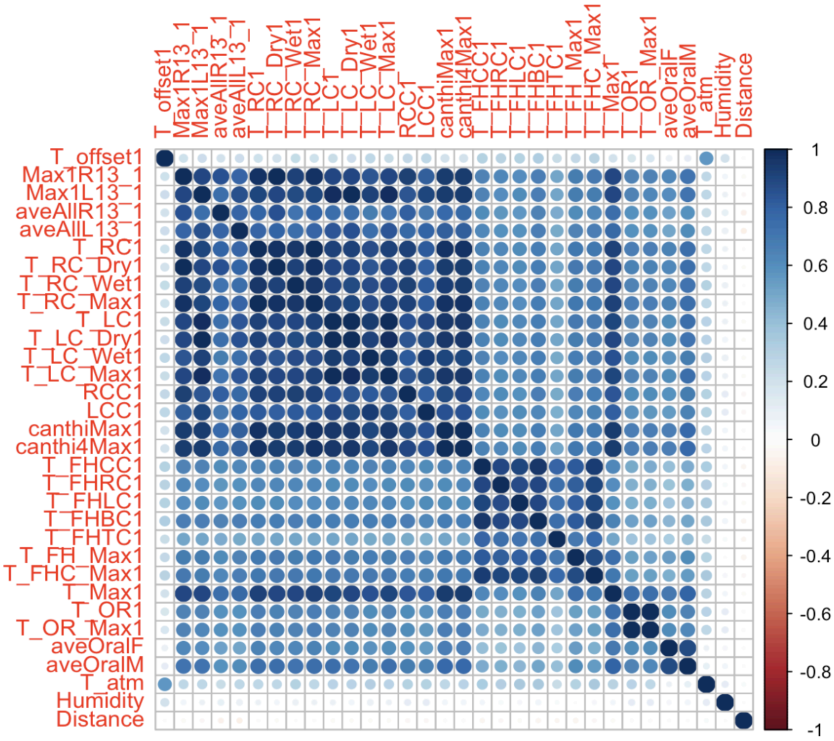

Figure 1. Correlation plot for continuous variables.

2 Performance of Tree-based Regression Models

2.1 Bagging

We trained the bagging model by using the randomForest function from the R package randomForest [17]. We set *mtry* parameter as 33, because bagging is just a random forest with m=p. Figure 2 shows the mean decrease of accuracy in predictions on the OOB samples for aveOralF and aveOralM. The higher values of mean decrease in accuracy indicate which predictors are more important to the model. From Figure 2, T_OR1 and T_Max1 are the most important predictors for both targets, aveOralF and aveOralM, and T_OR_Max1, aveAllL13_1, and T_atm also have strong effects on both targets. On the contrary, gender, T_FH_Max1, and aveAllR13_1 are only influential on aveOralM, and T_FHLC1, T_FHRC1, and T_FHC_Max1 are only influential on aveOralF. This suggests that a separate modeling is necessary for those two responses.

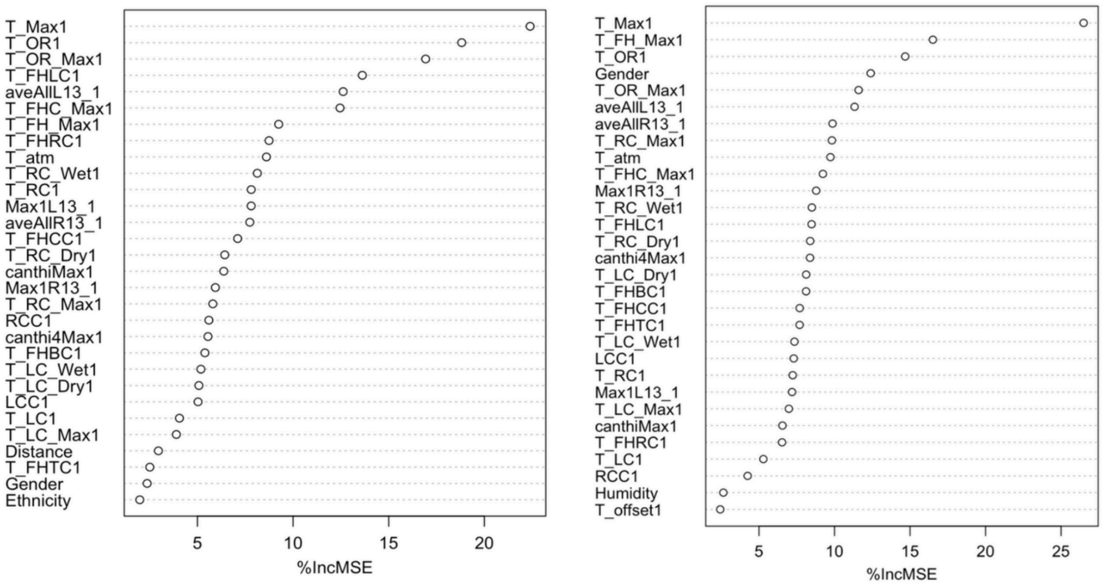

Figure 2. The mean decrease of accuracy in predictions on the OOB samples for each target. The left plot is for aveOralF. The right plot is for aveOralM.

2.2 Random Forest

The random forest model was trained by using the randomForest function from randomForest package [17]. We used the default number of trees as 500 and tuned the *mtry* parameter by using rfcv function and performing 5-fold cross-validation. In both training models, setting *mtry* as 16 has the best performance. Figure 3 shows the mean decrease of accuracy in predictions on the OOB samples for aveOralF and aveOralM. Similar to the results of the bagging model, the RF shows that T_OR_Max1, T_OR1, and T_Max1 are the top three most important predictors for both responses. T_FHLC1 and aveAllL13_1 have stronger effects on aveOralF, gender and T_FH_Max1 have stronger effects on aveOralM respectively.

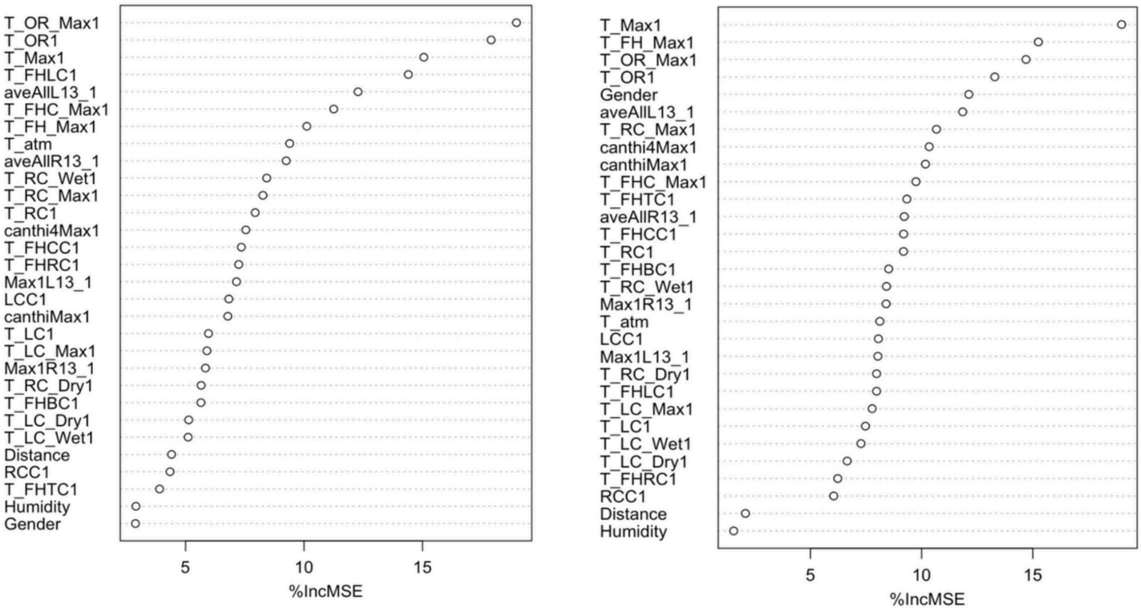

Figure 3. The mean decrease of accuracy in predictions on the OOB samples for each target. The left plot is for aveOralF. The right plot is for aveOralM

2.3 Boosting

For Boosting, we used the wbart function from the R-package gbm [18]. Performing cross-validation to obtain the best model, a total of 5000 trees were finally used the depth of each tree is 4, and the shrinkage parameter is 0.001. Figure 4 shows the relative influence of each variable on the training set. Figure 5 displays the partial dependence plot of boosting, which illustrates the marginal effect of the selected variables on the response after integrating the other variables.

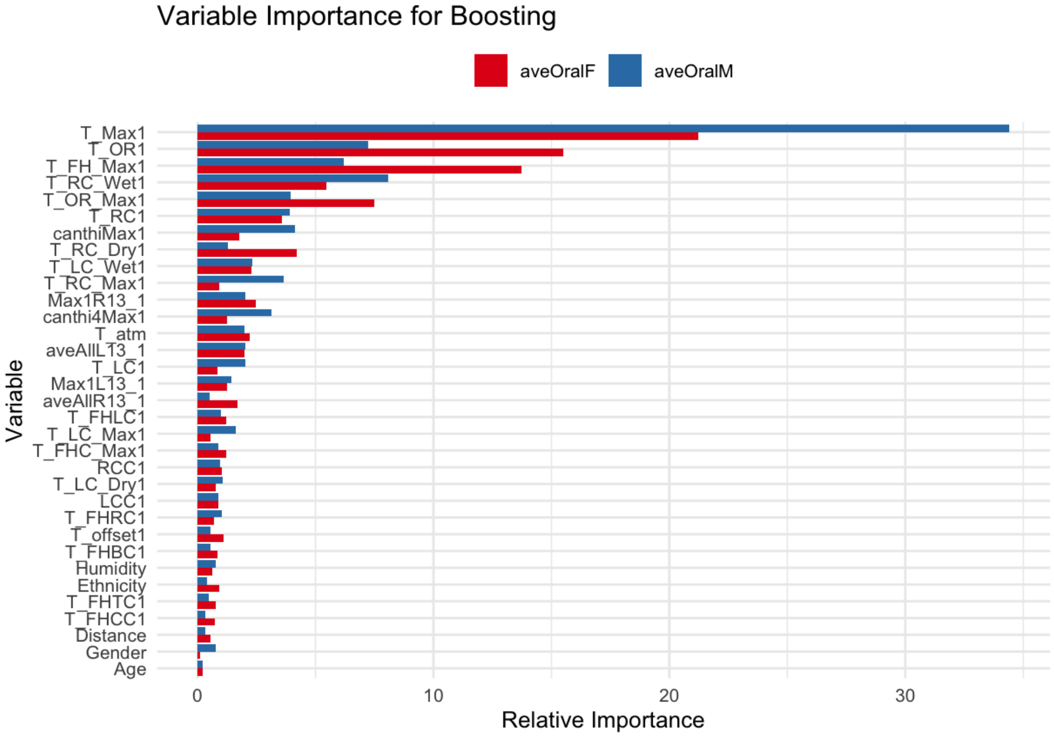

Figure 4. The relative influence of each variable on the training set. Red bars represent aveOralF, and the blue bars represent aveOralM

From Figure 4, we can see that T_Max1 emerges as the most significant predictor for both, with a higher impact on aveOralM. The influence of T_OR1 is substantial on aveOralF yet markedly less on aveOralM, while predictors like T_RC_Dry1, aveAllL13_1, and T_LC_Wet1 are influential only for aveOralF. Conversely, canthiMax1, canthi4Max1, T_RC1, and Max1L13_1 exhibit stronger effects on aveOralM, indicating that each response is governed by a unique set of dynamics. This suggests that separate models may be more suitable for effectively predicting each response, tailored to their respective influential predictors.

Figure 5 shows the partial dependence plots for T_Max, which is the most significant predictor for both aveOralF and aveOralM. While aveOralF shows a gradual increase, aveOralM exhibits a sharp rise after the T_Max1 value of approximately 36.5, indicating a potential threshold effect. The plot suggests a non-linear relationship and possible interactions with other variables, highlighting the complex dynamics T_Max1 has with the responses.

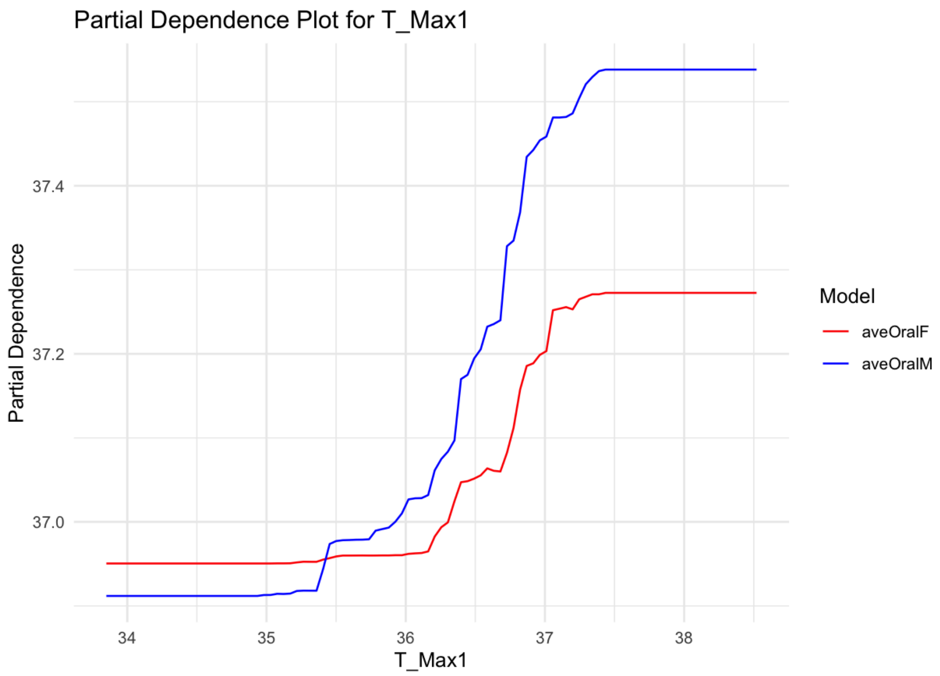

Figure 5. The partial dependence plots for T_Max1 on the training set. Red lines represent aveOralF, and the blue lines represent aveOralM

2.4 Bayesian Additive Regression Trees

For BART, we used the wbart function from the R-package BART [19]. A total of 200 trees were used, and the number of cross-validation folds to perform was 5. Figure 6 shows the percentages of each variable used in the BART models, which represents the importance of each variable.

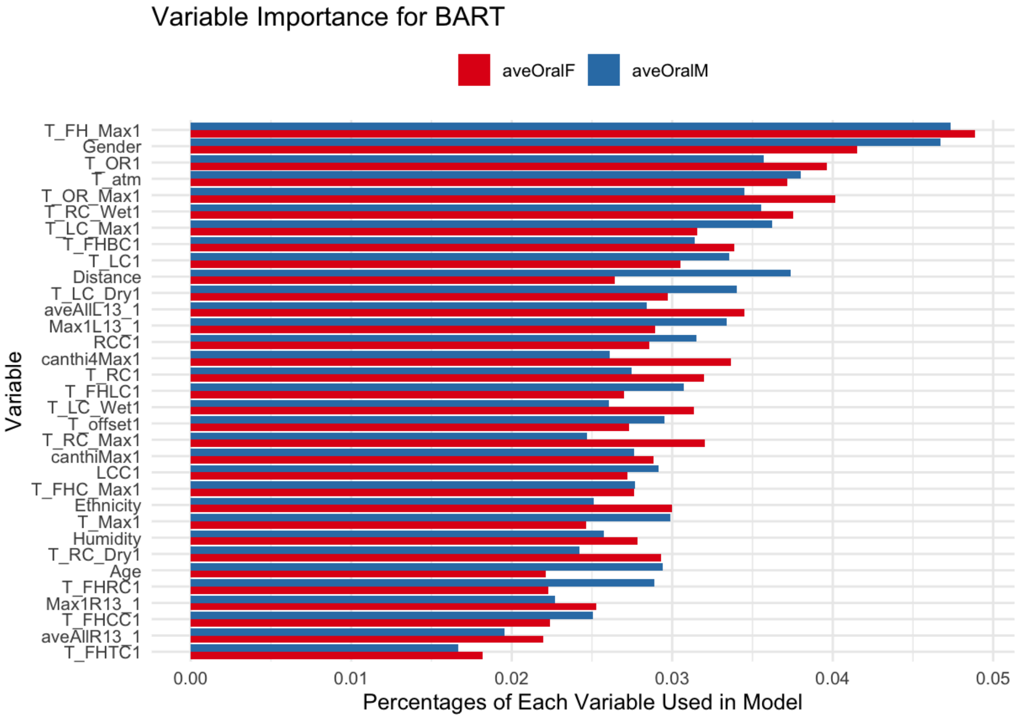

Figure 6. The percentages of each variable used in the BART model on the training set. Red bars represent aveOralF, and the blue bars represent aveOralM

From Figure 6, we can see that variables at the top of the plot, such as T_RC1 and T_RC_Wet1, are more influential in the model for aveOralM, whereas canthiMax1 shows greater importance for aveOralF. Several variables display a similar level of importance across both models, indicating their consistent impact on both responses.

3 Comparison and Best Model

We performed cross-validation and calculated the root mean square error (RMSE) for each model. Table 3 shows the performances of four thee-based regression methods on the test set. Based on the RMSE values provided in Table 3, the boosting method outperforms Bagging, RF, and BART for both aveOralF and aveOralM responses, with the lowest RMSE scores of 0.046 and 0.063 respectively. Therefore, boosting is the best model among those tested for this dataset, according to RMSE as a performance metric.

Table 3. The performances of four thee-based regression methods on the test set.

	Bagging	RF	**Boosting**	BART
RMSE for aveOralF	0.234	0.234	**0.046**	0.129
RMSE for aveOralM	0.259	0.259	**0.063**	0.195

Compared to Chayabhan's analysis, which uses facial feature temperatures from IR images, our model includes demographic data and temperatures from the entire body. They used a 90% training and 10% testing split, unlike ours. Despite these differences, comparing the RMSE values of both models is informative. Figure 7 shows the identification of the best regression model for each IR thermograph input feature, frame, and pair using RMSE from Chayabhan's. We can see that Chayabhan's best-performing linear regression model had an RMSE of 0.353, while our tree-based boosting model achieved a superior RMSE of 0.046, indicating that our model is more effective.

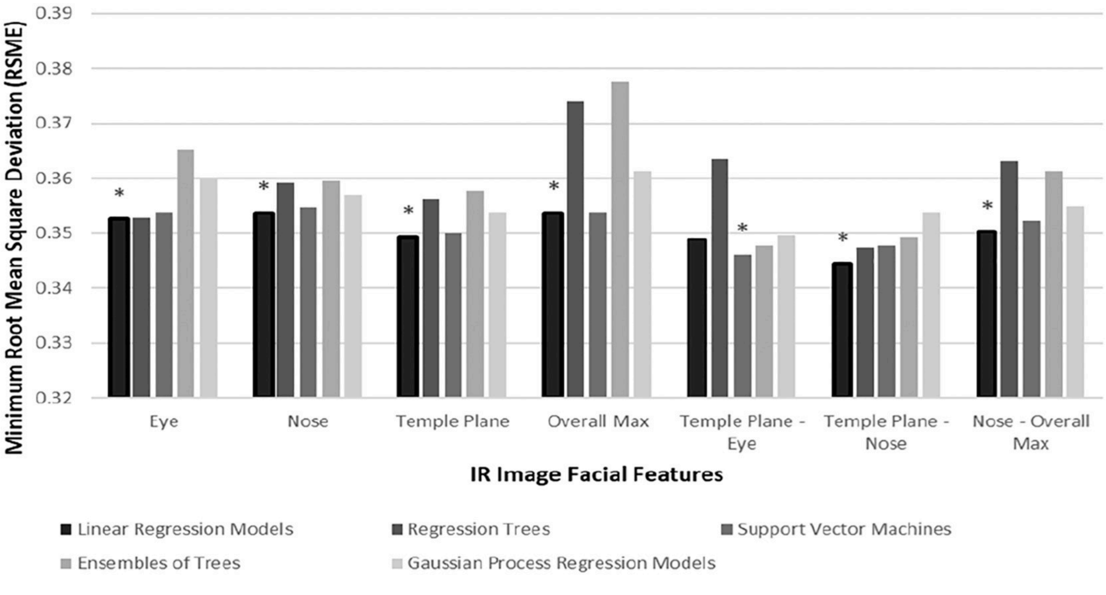

Figure 7. Identification of the best regression model for each IR thermograph input feature, frame, and pair using RMSE from Chayabhan's

Conclusions and Future Work

The investigation into the performance of tree-based regression methods on the given dataset yielded insightful results. The bagging, RF, and BART methods demonstrated variable importance and prediction accuracy, as evidenced by the mean decrease in accuracy and the usage percentages in the respective models. Across these methods, certain predictors like T_Max1, T_OR1, and canthiMax1 were consistently influential for the responses aveOralF and aveOralM. However, the boosting model, utilizing a granular approach with 5000 trees, a depth of 4, and a shrinkage parameter of 0.001, significantly outshined the other models in terms of RMSE, making it the superior method for this dataset.

The superior performance of the boosting model sets a promising direction for future research. The next steps could involve a deeper analysis of the threshold effects observed in the boosting model, particularly for the T_Max1 variable, which may inform more nuanced predictive modeling. Expanding the feature space with additional predictors, experimenting with other forms of boosting algorithms, and applying the model to different datasets to validate its robustness and versatility are logical extensions of this work. Furthermore, exploring the integration of the identified influential predictors into a more complex ensemble model that can perhaps combine the strengths of boosting and BART could also be beneficial. Finally, examining the applicability of these methods in different domain-specific contexts would further solidify their utility and enhance their generalizability across varied predictive modeling scenarios.

Reference

[1] Chiu W, Lin P, Chiou HY, et al. Infrared Thermography to Mass-Screen Suspected Sars Patients with Fever. *Asia Pacific Journal of Public Health*. 2005;17(1):26-28.

[2] Huang, C.; Wang, Y.; Li, X.; Ren, L.; Zhao, J.; Hu, Y.; Zhang, L.; Fan, G.; Xu, J.; Gu, X. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. Lancet 2020, 395, 497–506.

[3] Zissis G.J., Wolfe W.L. The Infrared Handbook. Technical report, DTIC document. 1978

[4] Wang, Q., Zhou, Y., Ghassemi, P., McBride, D., Casamento, J.P., & Pfefer, T.J. (2021). Infrared Thermography for Measuring Elevated Body Temperature: Clinical Accuracy, Calibration, and Evaluation. *Sensors (Basel, Switzerland), 22*

[5] Ismael Fernández-Cuevas, Joao Carlos Bouzas Marins, Javier Arnáiz Lastras, Pedro María Gómez Carmona, Sergio Piñonosa Cano, Miguel Ángel García-Concepción, Manuel Sillero-Quintana, Classification of factors influencing the use of infrared thermography in humans: A review, Infrared Physics & Technology, Volume 71, 2015

[6] Hastie T, Tibshirani R, Friedman J. The elements of statistical learning. Springer series in statistics[J]. New York, NY, USA, 2001.

[7] Breiman L. Random forests[J]. Machine learning, 2001, 45: 5-32.

[8] Chayabhan Limpabandhu, Frances Sophie Woodley Hooper, Rui Li, Zion Tse, Regression model for predicting core body temperature in infrared thermal mass screening, IPEM-Translation, Volumes 3–4, 2022, 100006, ISSN 2667-2588, https://doi.org/10.1016/j.ipemt.2022.100006.

[9] Wang, Q., Zhou, Y., Ghassemi, P., Chenna, D., Chen, M., Casamento, J., Pfefer, J., & Mcbride, D. (2023). Facial and oral temperature data from a large set of human subject volunteers (version 1.0.0). *PhysioNet*.

[10] Wikipedia contributors. "Bootstrap aggregating." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 3 Feb. 2024. Web. 21 Mar. 2024.

[11] Breiman, L. Bagging predictors. *Mach Learn* **24**, 123–140 (1996).

[12] Wikipedia contributors. "Random forest." *Wikipedia, The Free Encyclopedia*. Wikipedia, The Free Encyclopedia, 6 Mar. 2024. Web. 22 Mar. 2024.

[13] Friedman J H. Greedy function approximation: a gradient boosting machine[J]. Annals of Statistics, 2001: 1189-1232.

[14] Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)[J]. The annals of statistics, 2000, 28(2): 337-407.

[15] Chipman H A, George E I, McCulloch R E. BART: Bayesian additive regression trees[J]. 2010.

[16] Hastie T, Tibshirani R. Bayesian backfitting (with comments and a rejoinder by the authors[J]. Statistical Science, 2000, 15(3): 196-223.

[17] Liaw A, Wiener M (2002). “Classification and Regression by randomForest.” *R News*, **2**(3), 18-22.

[18] Ridgeway G. Generalized Boosted Models: A guide to the gbm package[J]. Update, 2007, 1(1): 2007.

[19] Sparapani R, Spanbauer C, McCulloch R. Nonparametric machine learning and efficient computation with Bayesian additive regression trees: the BART R package[J]. Journal of Statistical Software, 2021, 97: 1-66.