Journal of Probability and Statistical Science
24(2), 1-15Aug. 2026

Predicting Oral Temperature from Infrared Thermography and Environmental Data using Ensemble Machine Learning

Syed Muhammad Zeeshan
Department of Statistics
University of Peshawar, Pakistan

Muhammad Osama
Department of Statistics
Quaid-i-Azam University, Pakistan

Muhammad Salman
Department of Statistics
Quaid-i-Azam University, Pakistan

ABSTRACT

Infrared thermography (IRT) is a non-contact imaging technology that enables the assessment of surface temperature patterns and has potential applications in healthcare, including temperature monitoring and physiological assessment. However, the relationship between facial skin temperature and oral temperature is influenced by multiple physiological, environmental, and measurement-related factors, making accurate prediction a challenging regression problem. This study presents a comparative evaluation of four machine learning regression algorithms—Random Forest, K-Nearest Neighbors (KNN), Light Gradient Boosting Machine (LightGBM), and Extreme Gradient Boosting (XGboost) for predicting oral temperature from infrared thermography data. The study utilized a dataset containing 816 observations and 33 predictive features representing demographic characteristics, environmental conditions, and thermographic measurements obtained from facial regions. Model performance was assessed using the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and coefficient of determination (R^2). Among the evaluated approaches, the Random Forest Regressor achieved the strongest overall performance, obtaining an R^2 of 0.5608, an RMSE of 0.2299, and an MAE of 0.1722. LightGBM produced a comparable MAE of 0.1721 and an R^2 of 0.5476, whereas KNN and XGBoost achieved lower predictive performance. Feature importance analysis indicated that thermographic measurements associated with the periocular and facial regions were among the most influential predictors of oral temperature. The findings demonstrate the potential of machine learning-assisted infrared thermography for non-contact temperature estimation while also indicating that substantial unexplained variability remains. Future research should investigate systematic hyperparameter optimization, robust feature selection, external validation, larger and more diverse datasets, and clinically relevant validation protocols before deployment in real-world healthcare environments.

Keywords: Infrared thermography; machine learning; oral temperature prediction; Random Forest; KNN; LightGBM; XGBoost; medical thermography; non-contact temperature measurement.

1. Introduction

1.1 Background

Infrared thermography (IRT) is a non-invasive and non-contact imaging modality that measures infrared radiation emitted from an object's surface and converts this radiation into temperature-related information. In healthcare, infrared thermography has attracted considerable interest because it can provide rapid thermal measurements without requiring direct physical contact with the subject. This characteristic makes the technology potentially

- Received July 27, 2026, in final form August 2026.
- Muhammad Osama (corresponding author) is affiliated with Department of Statistics, Quaid-i-Azam University, Pakistan
  Email: muhammadosama0846@gmail.com

useful for applications involving temperature monitoring, physiological assessment, and medical screening.

Human skin temperature is influenced by several physiological and environmental factors, including blood circulation, metabolic activity, ambient temperature, humidity, and measurement distance. Consequently, surface temperature measurements obtained through infrared thermography cannot always be directly interpreted as equivalent to core body temperature. The development of computational models capable of learning the complex relationship between thermographic measurements and reference body temperature may therefore improve the utility of IRT for non-contact temperature estimation.

Machine learning provides a promising framework for this problem because thermographic datasets typically contain multiple interacting variables and potentially non-linear relationships. Unlike conventional linear regression approaches, tree-based ensemble algorithms and non-parametric methods can model complex relationships without requiring the explicit specification of interaction terms or a predetermined functional form.

The present study investigates the performance of four machine learning regression algorithms—Random Forest, K-Nearest Neighbors (KNN), LightGBM, and XGBoost—for oral temperature prediction using infrared thermography data. The analysis focuses on comparing predictive performance and identifying the thermographic characteristics that contribute most strongly to model predictions.

1.2 Problem Statement

Estimating oral or core temperature from facial surface temperature measurements is a complex regression problem. Although infrared thermography offers a rapid and contactless measurement mechanism, the observed thermal patterns are affected by environmental conditions, individual characteristics, anatomical location, and acquisition procedures.

The challenge is therefore not simply to measure temperature but to establish a reliable predictive relationship between multiple thermographic and contextual variables and a reference oral temperature measurement. Traditional statistical models may have limited ability to capture non-linear interactions among these variables. Machine learning models may provide a more flexible alternative; however, their performance can vary substantially depending on the characteristics of the dataset and the underlying relationships between predictors and the target variable.

A systematic comparison of different regression algorithms is consequently necessary to determine which modelling approaches are most appropriate for this prediction task.

1.3 Research Objectives

The primary objectives of this study are:

1. To develop machine learning regression models for predicting oral temperature using infrared thermography data.
2. To compare the predictive performance of Random Forest, KNN, LightGBM, and XGBoost.
3. To evaluate model performance using MAE, MSE, RMSE, and R^2.
4. To identify the most influential thermographic and contextual predictors using feature importance analysis.
5. To investigate relationships and potential redundancy among the predictor variables.
6. To assess the potential and limitations of machine learning-based infrared thermography for non-contact temperature estimation.

1.4 Research Contributions

This study makes several contributions to the field of machine learning-assisted medical thermography. First, it provides a comparative evaluation of four widely used regression approaches under a common experimental setting. Second, it investigates the relative contribution of thermographic features to oral temperature prediction. Third, it demonstrates the potential of ensemble learning methods for modelling non-linear relationships in infrared thermography data. Finally, the study identifies methodological limitations and future research requirements that should be addressed before such systems can be considered for practical healthcare applications.

2. Related Work

Machine learning has become an important component of modern healthcare research because of its ability to identify complex relationships within high-dimensional clinical and physiological datasets. Previous studies have demonstrated the effectiveness of machine learning techniques for disease prediction, medical diagnosis, and physiological parameter estimation [1–4]. In particular, ensemble learning algorithms such as Random Forest have been widely adopted for structured medical datasets because of their ability to model non-linear relationships, capture interactions among predictors, and provide estimates of feature importance [5]. Gradient boosting methods have also demonstrated strong predictive capabilities in healthcare applications, with XGBoost providing an efficient regularised framework for structured data analysis [6]. Similarly, LightGBM was introduced as an efficient gradient boosting decision-tree framework capable of handling large-scale datasets while maintaining competitive predictive performance [7]. K-Nearest Neighbors (KNN) has also been used in biomedical prediction and classification because of its non-parametric nature and ability to model local patterns in feature space [8]. Alongside conventional machine learning, infrared thermography (IRT) has emerged as a promising non-invasive and non-contact technology for medical and physiological assessment. Ring and Ammer [9] provided a comprehensive overview of infrared thermal imaging in medicine, highlighting its potential applications across different clinical domains. Lahiri et al. [10] similarly reviewed medical applications of infrared thermography and discussed its use for detecting physiological changes associated with disease. Research has investigated thermography for conditions including diabetes and vascular abnormalities [11,12], while other studies have explored its application in breast cancer screening and diagnosis [13,14]. The use of thermography for fever and temperature-related assessment has also received considerable attention. During infectious disease outbreaks, infrared thermal imaging was increasingly investigated as a non-contact approach for fever screening and temperature monitoring [15,16]. However, the reliability of thermographic temperature measurement can be affected by environmental temperature, measurement distance, camera characteristics, emissivity, and calibration conditions [17]. Recent studies have therefore investigated computational and machine learning methods to improve the robustness of thermal imaging systems. For example, machine learning combined with infrared thermography and computer vision has been used to predict human thermal states from facial and hand temperature measurements, with Random Forest models demonstrating strong predictive performance [18]. More recent research has proposed deep learning-based approaches for robust remote fever screening by explicitly addressing variations in measurement distance and environmental temperature [19]. In addition, recent reviews have highlighted the growing integration of artificial intelligence and infrared thermal imaging for medical diagnostics and physiological analysis [20]. These developments demonstrate that the combination of infrared thermography and machine learning represents a promising research direction for non-contact physiological assessment. Nevertheless, the prediction of oral

temperature from facial thermographic measurements remains challenging because surface temperature is influenced by multiple physiological, environmental, and measurement-related factors. Therefore, comparative evaluation of different machine learning regression algorithms is necessary to identify suitable predictive approaches and to determine which thermographic measurements provide the greatest predictive value. The present study addresses this research gap by comparing Random Forest, K-Nearest Neighbors, LightGBM, and XGboost for oral temperature prediction using infrared thermography data.

Although machine learning has been widely applied to medical thermography, the comparative performance of different regression algorithms for predicting oral temperature from facial thermographic measurements requires further investigation. In particular, it is important to determine whether complex boosting algorithms consistently outperform simpler ensemble approaches and which thermographic regions provide the greatest predictive value. This study addresses this gap by comparing four regression algorithms using the same dataset, preprocessing strategy, and evaluation metrics.

3. Materials and Methods

3.1 Dataset

The study utilized the Infrared Thermography Temperature dataset obtained from the UCI Machine Learning Repository. The dataset contains 816 observations and includes demographic characteristics, environmental variables, thermographic measurements, and oral temperature measurements.

The predictive variables consist of three broad categories:

- Demographic characteristics;
- Environmental and acquisition-related measurements; and
- Thermographic measurements obtained from multiple facial regions.

The target variable represents oral temperature measurements used as the reference for model training and evaluation.

The dataset contains no missing values according to the variable information available in the analysed dataset. The combination of multiple thermographic measurements with environmental and demographic information provides a multidimensional representation of the factors potentially associated with oral temperature.

3.2 Data Preprocessing

3.2.1 Categorical Data Encoding

Categorical variables were converted into numerical representations using one-hot encoding. This transformation allows machine learning algorithms that require numerical input to process categorical information while avoiding the assumption of an artificial ordinal relationship between categories.

3.2.2 Feature Scaling

Numerical variables were standardized using the standard scaler approach. The scaling transformation was fitted using the training data and subsequently applied to the test data. This procedure ensures that information from the test set does not influence the preprocessing parameters used during model development.

Feature scaling is particularly important for distance-based algorithms such as KNN, where variables with larger numerical magnitudes can disproportionately influence the distance calculation.

3.2.3 Training and Testing Strategy

The dataset was divided into training and testing subsets using an 80:20 split, with a random state of 42 to improve reproducibility. The training subset was used for model development, while the independent test subset was used to assess predictive performance.

Because the task involves continuous temperature prediction, evaluation should be interpreted as a regression problem rather than a classification problem. Therefore, performance was assessed using continuous regression metrics.

4. Machine Learning Models

4.1 Random Forest Regressor

Random Forest is an ensemble learning algorithm that constructs multiple decision trees using bootstrap samples of the training data. At each tree node, a random subset of features is considered for splitting. The final prediction is obtained by aggregating predictions from the individual trees. The method is well suited to structured datasets because it can capture non-linear relationships and interactions while generally being less sensitive to individual noisy observations than a single decision tree. Random Forest also provides an impurity-based measure of feature importance, which was used in this study to investigate the relative contribution of predictors.

4.2 K-Nearest Neighbors Regressor

KNN regression predicts the target value of an observation based on the target values of its nearest neighbors in feature space. In this study, an initial neighborhood size of five was used, with Euclidean distance and uniform weighting. The method provides a useful non-parametric baseline because it does not explicitly learn a global functional relationship. However, its performance may decline when the feature space becomes high-dimensional or when irrelevant and redundant variables affect the distance calculations.

4.3 LightGBM

LightGBM is a gradient boosting framework designed for efficient learning from structured data. It employs histogram-based learning and leaf-wise tree growth, enabling efficient model training while maintaining strong predictive performance.

Its performance on structured datasets has made it a widely used algorithm in machine learning competitions and practical applications. However, the effectiveness of LightGBM depends on appropriate hyperparameter configuration, including tree complexity, learning rate, and regularization.

4.4 XGBoost

XGboost is an optimized gradient boosting framework that sequentially constructs decision trees to minimize a specified loss function. The algorithm incorporates regularization mechanisms designed to reduce overfitting and provides efficient computational implementations.

Despite its strong reputation for tabular prediction tasks, its performance depends on appropriate hyperparameter selection. In the present analysis, the reported results were obtained under the experimental configuration used in the original analysis.

5. Evaluation Metrics

Four standard regression metrics were used to evaluate model performance.

I. Mean Absolute Error

Mean Absolute error is define as

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (1)$$

MAE measures the average absolute difference between observed and predicted values. Lower values indicate better predictive performance.

II. Mean Squared Error

Mean squared error is define as

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

MSE assigns greater penalties to larger prediction errors because the differences are squared.

III. Root Mean Squared Error

Root mean squared error is define as

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \quad (3)$$

RMSE represents the typical magnitude of prediction error in the same unit as the target variable. Lower RMSE values indicate better performance.

IV. Coefficient of Determination

Coefficient of determination is denoted by R^2.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)}{\sum_{i=1}^{n}(y_i - \bar{y})} \quad (4)$$

R^2 represents the proportion of variance in the target variable explained by the predictive model. Higher values indicate greater explanatory performance.

6. Results

6.1 Comparative Model Performance

The performance of the four evaluated models is presented in Table 1.

Table 1. Comparative Performance of Machine Learning Regression Models

Model	MAE	MSE	RMSE	R^2
Random Forest	**0.1722**	**0.0529**	**0.2299**	**0.5608**
LightGBM	0.1721	0.0544	0.2333	0.5476
KNN	0.1816	0.0592	0.2432	0.5085
XGboost	0.1802	0.0596	0.2442	0.5047

The Random Forest model achieved the highest R^2 value of 0.5608, indicating that it explained approximately 56.08% of the variance in the test-set target values under the experimental conditions used in this study. It also achieved the lowest RMSE at 0.2299. LightGBM demonstrated closely comparable performance, obtaining an R^2 of 0.5476 and an RMSE of 0.2333. Interestingly, LightGBM achieved the lowest MAE at 0.1721, marginally outperforming Random Forest at 0.1722. KNN achieved an R^2 of 0.5085, while XGBoost obtained an R^2 of 0.5047. Both models demonstrated higher RMSE values than the two leading ensemble approaches. Overall, the results indicate that Random Forest provided the strongest

aggregate performance according to R^2 and RMSE, while LightGBM produced a marginally lower average absolute error.

6.2 Actual Versus Predicted Values

The actual-versus-predicted analyses indicated that Random Forest and LightGBM produced predictions that were more closely aligned with the ideal identity relationship than KNN and XGboost. The greater concentration of observations around the ideal prediction line for Random Forest suggests comparatively better agreement between predicted and observed temperatures. However, the dispersion of predictions also indicates that the models do not fully capture the variability present in the target variable. This observation is consistent with the moderate R^2 values obtained by all four algorithms. Although machine learning models captured a meaningful proportion of the predictive signal, approximately 44% of the variance remained unexplained by the best-performing model in the evaluated test set.

6.3 Feature Importance

Feature importance analysis based on the Random Forest model identified thermographic measurements from the periocular and facial regions as the most influential predictors. The most important predictors were primarily associated with average and maximum temperature characteristics of facial regions, particularly measurements obtained around the eye area. Additional important variables included temperature measurements from the canthi, forehead, and cheek regions. These results suggest that localized facial thermal patterns may provide valuable information for estimating oral temperature. However, feature importance should not be interpreted as evidence of causality. Rather, it indicates the relative contribution of predictors to the fitted Random Forest model.

6.4 Correlation Analysis

The correlation analysis revealed strong relationships among several thermographic variables obtained from similar anatomical regions. Some predictor pairs exhibited correlation coefficients greater than 0.80, suggesting substantial redundancy. Such correlations are expected because measurements obtained from neighboring regions may reflect similar underlying thermal processes. Although tree-based models can generally tolerate correlated variables, redundant predictors may influence feature importance estimates and can complicate the interpretation of individual variables. Environmental factors also demonstrated relationships with selected thermographic measurements, supporting the importance of controlling environmental conditions during infrared temperature acquisition.

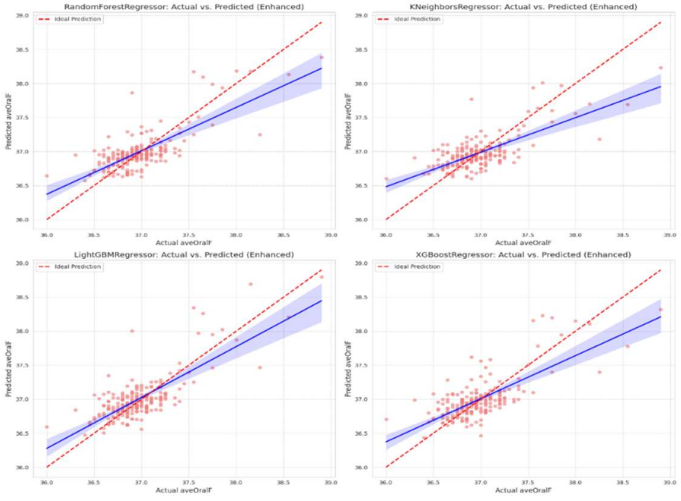


Figure 1. Shows Actual Vs Predicted plots for adopted machine learning models

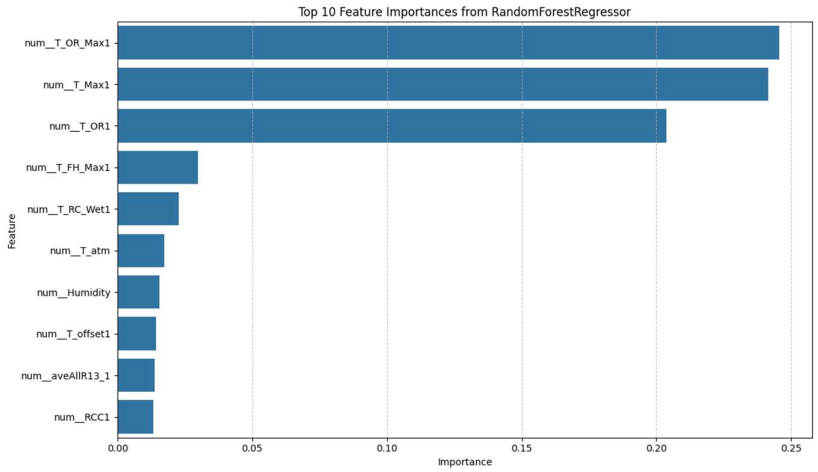


Figure 2. Shows important features from Random Forest

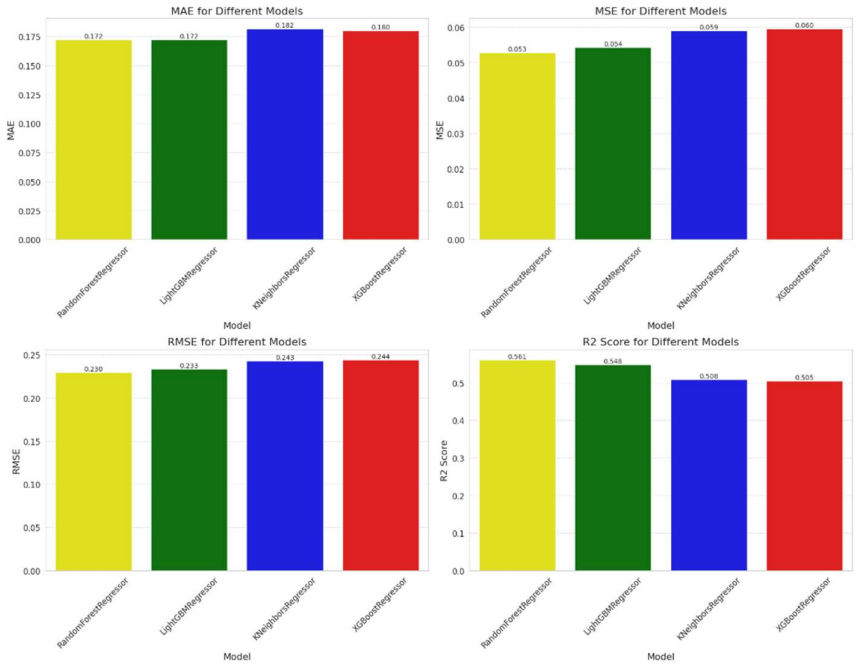


Figure 3. Shows Models predictive performance

7. Discussion

7.1 Interpretation of Model Performance

The results demonstrate that Random Forest achieved the strongest overall predictive performance among the four evaluated algorithms. Its superior R^2 and RMSE values suggest that the ensemble of decision trees was effective in modelling the non-linear relationships between thermographic measurements and oral temperature. The performance of Random Forest may be explained by its ability to capture complex interactions without requiring the researcher to explicitly define these relationships. Thermographic data often contain interdependent variables originating from different anatomical regions, and the relationship between these measurements and oral temperature may not be adequately represented by a

simple linear function. LightGBM also demonstrated strong predictive performance, achieving a slightly lower R^2 but the lowest MAE. This indicates that its typical absolute prediction error was marginally lower, although its larger squared-error contribution resulted in a slightly higher RMSE. KNN performed less effectively than the two leading ensemble approaches. One possible explanation is that the relationship between thermographic observations and oral temperature is not sufficiently local in the scaled feature space. In addition, redundant variables and high-dimensional feature representations can reduce the effectiveness of distance-based approaches. XGboost produced the lowest R^2 among the four models. This result does not necessarily imply that XGBoost is inherently unsuitable for thermographic temperature prediction. Rather, its performance may be influenced by the selected hyperparameters and experimental configuration. More systematic optimisation may produce different results.

7.2 Importance of Thermographic Features

The feature importance analysis suggests that facial temperature measurements, particularly those associated with the periocular region, contain substantial predictive information regarding oral temperature. This finding is physiologically plausible because the regions surrounding the eyes are relatively vascularized and may exhibit thermal characteristics associated with systemic temperature regulation. Nevertheless, the results should be interpreted cautiously because thermal measurements are also influenced by ambient conditions, sensor characteristics, distance, and individual variability. The identification of multiple important features across different facial regions also suggests that oral temperature prediction may benefit from integrating information from several anatomical locations rather than relying on a single measurement.

7.3 Clinical Relevance

The findings demonstrate the potential of machine learning-assisted infrared thermography for non-contact temperature estimation. Such systems could be relevant in environments where rapid and contactless temperature assessment is desirable. However, the current predictive performance is not sufficient to establish clinical equivalence with conventional temperature measurement methods. The R^2 of 0.5608 indicates that the best-performing model explains only a moderate proportion of the observed variation in the test set. Therefore, the proposed approach should currently be considered a research-oriented predictive framework rather than a clinically validated replacement for established temperature measurement techniques.

7.4 Practical Implications

The results have several practical implications. First, the superior performance of Random Forest suggests that ensemble tree-based approaches may provide a useful baseline for structured infrared thermography datasets. Second, the feature importance results indicate that future data collection protocols may benefit from prioritizing informative facial regions. However, reducing the number of measurements should only be considered after systematic feature selection and independent validation. Third, environmental conditions should be carefully controlled or explicitly incorporated into predictive models. Variations in ambient temperature, humidity, and measurement distance may introduce systematic sources of error. Finally, future machine learning systems should prioritize both predictive accuracy and interpretability. In medical applications, a model's ability to explain its predictions is often as important as its raw predictive performance.

8. Conclusion

This study presented a comparative evaluation of four machine learning regression algorithms for predicting oral temperature from infrared thermography data. Among the evaluated models, Random Forest demonstrated the strongest overall performance, achieving an R^2 of 0.5608 and an RMSE of 0.2299. LightGBM produced comparable performance and achieved the lowest MAE, while KNN and XGBoost demonstrated relatively lower predictive accuracy under the evaluated experimental conditions. Feature importance analysis indicated that thermographic measurements from the periocular and broader facial regions were among the most informative predictors. This finding supports the potential value of facial thermal patterns for estimating oral temperature and highlights the importance of selecting informative thermographic measurements. Nevertheless, the moderate predictive performance indicates that substantial variability remains unexplained. Consequently, machine learning-based infrared thermography should not yet be considered a direct replacement for clinically validated temperature measurement techniques. Instead, the results provide evidence supporting further investigation into non-contact temperature prediction.

Data Availability

The study utilized the Infrared Thermography Temperature dataset available through the UCI Machine Learning Repository. The experimental analysis was conducted using the dataset configuration and modelling procedures described in this manuscript. https://archive.ics.uci.edu/dataset/925/infrared+thermography+temperature+dataset

References

[1]. Al Husaini, M. A. S., Habaebi, M. H., Hameed, S. A., Islam, M. R., & Gunawan, T. S. (2020). A systematic review of breast cancer detection using thermography and neural networks. *IEEE Access*, 8, 208922–208937.

[2]. Altman, N. S. (1992). An introduction to kernel and nearest-neighbor nonparametric regression. *The American Statistician*, 46(3), 175–185.

[3]. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

[4]. Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.

[5]. Deo, R. C. (2015). Machine learning in medicine. *Circulation*, 132(20), 1920–1930.Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine*, 25, 44–56.

[6]. FeverNet: Enabling accurate and robust remote fever screening. *(2024)*. *Pattern Recognition*, 156, 110755.

[7]. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30, 3146–3154.

[8]. Lahiri, B. B., Bagavathiappan, S., Jayakumar, T., & Philip, J. (2012). Medical applications of infrared thermography: A review. *Infrared Physics & Technology*, 55(4), 221–235.

[9]. Mohamed, E. A., Rashed, E. A., Gaber, T., & Karam, O. (2022). Deep learning model for fully automated breast cancer detection system from thermograms. *PLOS ONE*, 17(1), e0262349.

[10]. Nishide, K., Nagase, T., Oba, M., et al. (2009). Ultrasonographic and thermographic screening for latent inflammation in diabetic foot callus. *Diabetes Research and Clinical Practice*, 85(3), 304–309.

[11]. Nowakowski, A. Z., & Kaczmarek, M. (2025). Artificial intelligence in IR thermal imaging and sensing for medical applications. *Sensors*, 25(3), 891.
[12]. Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, 380(14), 1347–1358.
[13]. Ring, E. F. J., & Ammer, K. (2012). Infrared thermal imaging in medicine. *Physiological Measurement*, 33(3), R33–R46.
[14]. *Smart detection of indoor occupant thermal state via infrared thermography, computer vision, and machine learning*. (2023). *Building and Environment*, 228, 109811.
[15]. Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, 19, 64.
[16]. Sivanandam, S., Anburajan, M., Venkatraman, B., Menaka, M., & Sharath, D. (2012). Medical thermography: A diagnostic approach for type 2 diabetes based on non-contact infrared thermal imaging. *Endocrine*, 42(2), 343–351.
[17]. Ring, E. F. J., & Ammer, K. (2012). Infrared thermal imaging in medicine. *Physiological Measurement*, 33(3), R33–R46.
[18]. Zhou, Y., et al. (2021). Reliability enhancement of non-contact fever screening technology (thermography, etc.) for quarantine inspection. *Measurement: Sensors*, 18, 100160.
[19]. Zhou, Y., et al. (2021). Reliability enhancement of non-contact fever screening technology: Considerations related to device performance, target parameters, and ambient environmental conditions. *Measurement: Sensors*, 18, 100160.