

Computational Methods for Core Body Temperature Estimation from Wearable Physiological Signals



Yuanzhe Zhao
St Cross College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Hilary term, 2026

Supervised by Professor Jeroen Bergmann

Abstract

Monitoring human core body temperature (CBT) is important during physical activity and exposure to hot environments, supporting applications in occupational health, sports and training for the prevention of exertional heat illness. However, direct measurement of CBT using conventional methods such as rectal or gastrointestinal measurement is invasive, while peripheral temperature measurements such as axillary, oral and tympanic temperature are convenient but unreliable during physical activities. These limitations motivate the development of non-invasive CBT estimation methods from wearable physiological signals.

This thesis focuses on estimating human CBT under wearable sensing constraints, primarily using heart rate measurements, with limited use of skin temperature. The research addresses several key challenges in this problem, including the non-linear relationship between heart rate and CBT across different physiological states, uncertainty in physiological state transitions, sensor noise, and limited interpretability.

This work makes three main contributions. First, body temperature measurement and CBT estimation methods were systematically reviewed. This included systematic reviews of non-contact body temperature measurement techniques and data-driven CBT estimation methods, together with a standardized benchmark evaluation of selected methods using two independent datasets. Second, heart-rate-only state-space models were developed for real-time CBT estimation, including (i) a multi-model extended Kalman filter (EKF) framework using variance-based fusion to implicitly capture state-dependent thermoregulatory dynamics; (ii) a probabilistic state inference approach that couples a Hidden Markov model (HMM) with a particle filter (PF) to explicitly represent physiological state uncertainty; (iii) a residual-compensated adaptive Kalman filter that integrates a lightweight neural network to correct systematic errors of the observation model. Third, a physics-guided hybrid framework was developed by combining a first-principles two-node thermoregulation model with neural networks to improve estimation accuracy, robustness, and physiological interpretability.

On the dataset, the best Root Mean Square Error (RMSE) decreased progressively from 0.38 °C for the multi-model Kalman filtering framework to 0.359 °C for the HMM–PF framework, 0.31 °C for the residual-compensated adaptive Kalman filter, and 0.27 °C for the physics-guided hybrid framework. This progressive reduction demonstrates that the primary aim of improving CBT estimation accuracy was achieved under the evaluated conditions. The residual-compensated adaptive Kalman filter approached the pragmatic research target of approximately 0.3 °C, while the physics-guided framework achieved an RMSE below this target under controlled physical-activity conditions. These results provide an overall indication of methodological progress.

In conclusion, this research demonstrates that introducing appropriate structural constraints, including state dependence, probabilistic inference, adaptive correction, and physiological priors, can progressively improve the accuracy and reliability of wearable CBT estimation during physical activity. The proposed frameworks provide complementary trade-offs among accuracy, robustness, computational cost, adaptability, and physiological interpretability, offering a foundation for the future development of personalized, robust, and deployable wearable heat strain monitoring systems.

Acknowledgements

First and foremost, I would like to express my sincere gratitude to my supervisor, Professor Jeroen Bergmann, for his invaluable guidance, encouragement, and continued support throughout my DPhil. I would also like to thank my co-supervisors, Dr David Holdsworth and Dr Paulo Savaget, for their helpful discussions, advice, and support during my doctoral studies.

My thanks also go to Professor Robert Carlisle and Professor John Allen for examining this thesis and for their thoughtful and constructive comments, which helped me improve the final version of this work.

I am deeply grateful to all the friends and colleagues who have supported me throughout these years, both in Oxford and elsewhere. I would especially like to thank, in alphabetical order by surname, Dr Leonardo de Almeida e Bueno, Dr Chen Yuan, Dr Ding Zhiyuan, Dr Han yu, Phoebe Haste, Dr Hou Xinyu, Elyse Levens, Li Weihao, Ren Hangyu, Dr Joseph Russell, Wang Peijue, Dr Zhang Xiao, and Dr Zhang Xiaodong. Thank you for the discussions, meals, travels, laughter, and many moments we have shared, and for being part of this important period of my life.

I owe my deepest gratitude to my parents for their unconditional love, support, patience, and encouragement throughout my education. They have always supported the choices I have made and given me the confidence to pursue my own path. None of this would have been possible without them.

Finally, and most importantly, I would like to thank my Honey, Mary, for always being by my side. Thank you for your love, patience, understanding, encouragement, and companionship through all the ups and downs of this journey. You have shared not only the happy moments, but also the uncertainty, frustration, and long days that came with completing a DPhil. I am incredibly grateful to have shared this chapter of my life with you, and I look forward to all the chapters still to come.

Contents

Acronyms	viii
1 Introduction	1
1.1 Motivation and background	1
1.1.1 Core body temperature and human thermoregulation	1
1.1.2 Heat stress and the need for core body temperature monitoring	3
1.1.3 Why core body temperature remains difficult to monitor in the field	4
1.1.4 From measurement to estimation under wearable constraints .	5
1.2 Research objectives	7
1.3 Contributions and organization of the thesis	8
2 Non-Contact Infrared Thermometers and Thermal Scanners for Human Body Temperature Monitoring: A Systematic Review	12
2.1 Introduction	13
2.2 Materials and Methods	15
2.2.1 Search Strategy	15
2.2.2 Study Selection Criteria	15
2.2.3 Data Extraction	16
2.2.4 Quality Assessment	17
2.2.5 Data Synthesis	18
2.3 Results	18
2.3.1 Characteristics of Included Studies	18
2.3.2 Methodological Quality of Included Studies	26
2.4 Discussion	28
2.5 Conclusion	37
3 Data-Driven Estimation of Core Body Temperature during Physical Activity under Heat Stress: A Systematic Review and Standardized Evaluation	39
3.1 Introduction	40
3.2 Methods	44
3.2.1 Systematic Review	44
3.2.2 Standardized Benchmarking	48
3.3 Results	54
3.3.1 Model categories and inputs	65
3.3.2 Experiment settings	67
3.3.3 Reported performance across studies	69
3.3.4 Benchmark performance across the eligible studies	70
3.4 Discussion	73

4	Core Body Temperature Estimation from Heart Rate via Multi-Model Kalman Filtering and Variance-Based Fusion	81
4.1	Introduction	82
4.2	Method	85
4.2.1	Datasets	85
4.2.2	Multi-state Kalman Filters	86
4.2.3	Model evaluation methods	92
4.3	Results	93
4.3.1	Evaluation on Dataset 1	93
4.3.2	Evaluation on Dataset 2	96
4.4	Discussion	101
4.5	Conclusion	103
5	Real-Time Estimation of Core Body Temperature for Heat Stress Monitoring in Hot Environments Using Wearable Heart Rate Sensors	105
5.1	Introduction	106
5.2	Method	110
5.2.1	Datasets	111
5.2.2	The State-Space Framework for CBT Estimation	111
5.2.3	State Inference and Fusion Mechanisms	113
5.2.4	Bayesian Filtering Algorithms	117
5.2.5	Experimental Protocol and Evaluation	121
5.3	Results	121
5.4	Discussion	127
5.5	Conclusion	132
6	Estimating Core Body Temperature from Heart Rate Using a Residual-Compensated Adaptive Kalman Filter	133
6.1	Introduction	134
6.2	Methods	137
6.2.1	Dataset	137
6.2.2	Extended Kalman Filter (EKF)	138
6.2.3	Residual-Compensated Adaptive Kalman Filter (RCAKF)	140
6.2.4	Evaluation Metrics	143
6.2.5	Comparison Methods	144
6.3	Results	146
6.4	Discussion	152
6.5	Conclusion	159
7	A Physics-Guided Hybrid Model for Core Body Temperature Estimation: Thermoregulation model with Residual Learning	160
7.1	Introduction	161
7.2	Methods	165
7.2.1	Two-Node Thermoregulation Model	165
7.2.2	Residual Compensation with Neural Networks	169
7.2.3	Dataset	173
7.2.4	Evaluation Metrics and Comparative Settings	173
7.3	Results	175

7.3.1	Overall Estimation Performance	175
7.4	Discussion	181
7.5	Conclusion	188
8	Conclusions and Future Work	190
8.1	Conclusions	190
8.2	Discussion	193
8.2.1	Limitations	195
8.3	Future Work	196
8.3.1	Expansion of Dataset Size and Diversity	196
8.3.2	Advanced Data-Driven Models	197
8.3.3	Personalization and Online Learning	198
8.3.4	Generalization to Cold Environments	199
8.3.5	Multimodal Sensing Optimization	199
8.3.6	Integrated Operational Early-Warning Systems	199
8.3.7	Extension to Other Applications	200
Appendices		
A	Data and Reproducibility	203
A.1	Dataset Availability	203
A.2	Chapter 2–3: Supplementary Study Characteristics	203
A.3	Chapter 4: PSSK and TCBK	204
A.4	Chapters 5–7: Algorithmic Descriptions	204
A.5	Ethical considerations	206
References		207

Acronyms

AUC Area Under the Curve	MAE Mean Absolute Error
BKFB Biphasic Kalman Filter-Based Model	MAKF Moving Average Kalman Filter
CBT Core Body Temperature	MD Mean Difference
CNN Convolutional Neural Network	NATA National Athletic Trainers' Association
DTW Dynamic Time Warping	NCIT Non-Contact Infrared Thermometer
ECG Electrocardiogram	PF Particle Filter
EHI Exertional Heat Illness	PIML Physics-Informed Machine Learning
EHS Exertional Heat Stroke	PINN Physics-Informed Neural Network
EKF Extended Kalman Filter	PPG Photoplethysmography
FLOPs Floating Point Operations	PSSK Physiological State-Specific Kalman Filter
GRU Gated Recurrent Unit	RCAKF Residual-Compensated Adaptive Kalman Filter
HMM Hidden Markov Model	RMSE Root Mean Square Error
HR Heart Rate	SD Standard Deviation
IGA Improved Genetic Algorithm	TCBK Trial Clustering-Based Kalman Filter
IRT Infrared Thermography	ZHF Zero-Heat-Flux
KF Kalman Filter	
LoA Limits of Agreement	
LSTM Long Short-Term Memory	

1

Introduction

1.1 Motivation and background

1.1.1 Core body temperature and human thermoregulation

To understand human body temperature, firstly we need to distinguish between two similar concepts: core temperature and shell temperature [1, 2]. This distinction is useful for understanding the physiological mechanisms of human thermoregulation and for interpreting temperature measurement results obtained in clinical and real-world settings.

As shown in Figure 1.1, core temperature refers to the temperature of the body's internal environment, mainly the temperature of vital organs in the cranial, thoracic, and abdominal cavities, such as the brain, heart, lungs, and other essential organs. In healthy individuals, the core temperature is typically maintained within a narrow range around 37 °C, although there are slight differences between individuals[2–4]. The stability of core temperature is critical for normal physiological function, as enzymatic reactions and metabolic processes are sensitive to temperature fluctuations. To prevent metabolic disturbances caused by core temperature fluctuations, the human body relies on a complex thermoregulatory control system involving the hypothalamus, autonomic nervous system, and endocrine responses to maintain core body temperature within an optimal range [5, 6].

Core body temperature is not a passive variable, but a physiological state maintained by coordinated control mechanisms. In cold environments, heat loss is reduced through vasoconstriction, while metabolic heat production is increased through

shivering. Conversely, in hot environments, heat loss is increased through vasodilation and sweating, thus limiting further increases in core body temperature [2, 7].

Different from core temperature, shell temperature refers to the temperature of the body's outer tissues, including the skin and subcutaneous layers. Shell temperature is not strictly regulated and is more susceptible to the ambient temperature, airflow, clothing conditions, and physical activities, typically ranging from approximately 30 °C to 40 °C [1, 8]. Therefore, shell temperature varies with location and time, and is greatly affected by external factors, making it difficult to stably reflect the true thermal state inside the human body.

However, the shell plays a key role in thermoregulation by acting as a thermal buffer between the body core and the environment. It is the primary interface for heat exchange through radiation, conduction, convection, and evaporation [2, 3, 7]. In cold environments, reduced peripheral blood flow lowers shell temperature and limits heat loss. In hot environments, increased skin blood flow elevates shell temperature and supports heat dissipation, particularly when combined with sweating.

The interaction between core and shell temperatures is therefore dynamic and continuous. Thermal information is detected by a distributed network of thermoreceptors: central thermoreceptors located in the hypothalamus and other deep body structures primarily sense changes in core body temperature, while peripheral thermoreceptors in the skin monitor shell temperature and environmental exposure [1, 8]. These signals are integrated by the hypothalamus, which orchestrates appropriate physiological responses to preserve thermal balance. For example, during physical activity, increased metabolic heat production leads to a rise in core body temperature, requiring enhanced skin perfusion and sweating to promote heat dissipation, thus raising the shell temperature [2, 3].

In conclusion, shell temperature reflects heat exchange with the ambient environment, while core body temperature represents the internal thermal state directly related to health and thermal risk. Recognizing this distinction is fundamental for understanding both the limitations of surface temperature measurements and the challenges in monitoring internal thermal strain under real-world and wearable-constrained conditions.

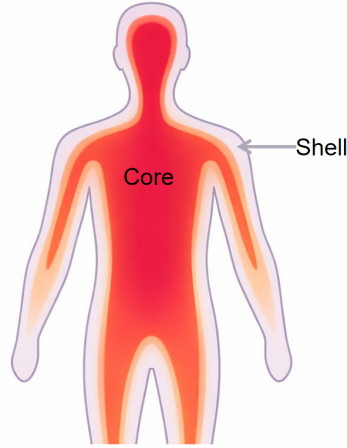


Figure 1.1: Illustration of core and shell temperature in thermoregulation.

1.1.2 Heat stress and the need for core body temperature monitoring

Core body temperature (CBT) is the primary physiological indicator of internal heat strain during physical activity and exposure to hot environments. In high-risk populations such as athletes, military personnel, and outdoor workers, sustained elevations in CBT are closely associated with impaired physical performance, reduced cognitive function, and increased risk of exertional heat illness (EHI), including the more dangerous exertional heat stroke (EHS) [9–13].

Exertional heat stroke is typically characterized by severe hyperthermia, with CBT often exceeding 40 °C, together with signs of organ-system dysfunction resulting from impaired thermoregulatory control [14, 15]. The risk of morbidity and mortality increases sharply as both the magnitude and duration of hyperthermia increase. In contrast, timely and effective cooling interventions can reduce the risk [16, 17]. These factors have led researchers and practitioners to pay increasing attention to continuous heat-strain monitoring to support individualized interventions, and work–rest strategies in sport, occupational, and military scenarios [18–20].

In practice, the environmental assessment is a simple and commonly used method for assessing thermal risks. The Wet-Bulb Globe Temperature (WBGT) is the most widely adopted metric. It integrates ambient temperature, humidity, solar radiation, and wind speed to provide a single indicator of external heat load [21, 22]. WBGT is useful for event and training scheduling by identifying environmental conditions

with high heat stress [23, 24]. Each of these environmental factors affects human heat exchange through a different mechanism. Air temperature determines the temperature gradient between the skin and the surrounding air and therefore mainly influences the convective heat transfer. Relative humidity affects evaporative heat loss: when humidity is high, the reduced water-vapour pressure gradient between the skin and the environment limits sweat evaporation and promotes heat storage. Solar radiation increases radiant heat gain, particularly during outdoor activity, while wind speed influences both convective and evaporative heat transfer. [24, 25] However, environmental indicators alone cannot capture the determinants of internal heat strain. CBT responses strongly rely on metabolic heat production (e.g. exercise intensity and exposure time), clothing conditions, personal fitness and acclimation status, as well as the individual thermoregulatory capacity. All of these factors influence the rate of heat accumulation and dissipation [25, 26]. As a result, heat-related illness can still occur under WBGT-guided management, motivating the need for physiological monitoring for tracking internal thermal strain in real time [27, 28].

1.1.3 Why core body temperature remains difficult to monitor in the field

Though core body temperature is important in assessing internal thermal strain, accurate monitoring of CBT is still a challenging question outside clinical environments. The primary difficulty is the trade-off between measurement accuracy and practical deployability. Traditional CBT measurements such as rectal and gastrointestinal temperature are commonly regarded as golden reference methods for CBT assessment during exercise [15, 20, 29, 30]. However, these methods are invasive and limited by discomfort, hygiene concerns, cost-effectiveness, and operational burden. Therefore, they are unsuitable for continuous monitoring during physical activity in practice [31–34].

In contrast, shell temperature measurements such as the oral, axillary, or forehead temperature are more convenient, but they do not reliably reflect CBT under dynamic conditions. Because the shell acts as a thermal buffer between the body core and the environment, its temperatures are easily influenced by airflow and sweat evaporation [2, 31, 35]. Tympanic temperature is commonly used in clinical practice as

a surrogate measure of core temperature. However, studies have shown that tympanic measurements may also have insufficient agreement with reference CBT measurements during physical activity [36, 37]. Therefore, both shell temperature measurements and tympanic temperature may show poor agreement with gold-standard CBT measurements during physical activity and environmental heat exposure, limiting their application for real-time heat-strain monitoring [38].

To achieve accurate and non-invasive core body temperature measurement, a number of hardware solutions have emerged, focusing on sensor design and sensing mechanisms. Ingestible telemetric capsules are commonly used in field studies and with high accuracy, but they are constrained by single-use consumables, high-cost, unstable gastrointestinal transit times, and reliance on dedicated receivers [29, 32]. Zero-heat-flux (ZHF) thermometry and heat-flux(HF) sensors provide promising non-invasive approaches with good agreement to reference CBT under certain conditions [39, 40]. Nevertheless, these technologies rely on proper sensor placement, stable thermal contact, and specialised hardware. Their performance may therefore be affected by vigorous movement, heavy sweating, or changes in environmental conditions [31, 40, 41].

Together, these limitations highlight a challenging trade-off between the accuracy of CBT measurement and the practical feasibility of direct measurement. As a result, developing accurate and non-invasive approaches to measure or estimate CBT using commonly available sensing technologies would offer substantial practical value for applications in sports, occupational health, and everyday life. In this context, estimating CBT using physiological signals from wearable devices has become a promising research direction.

1.1.4 From measurement to estimation under wearable constraints

From a physiological perspective, this estimation paradigm is based on the relationship between cardiovascular activity and the internal heat balance. During physical activity, heart rate increases primarily to meet the oxygen demand of working muscles, and is closely associated with metabolic heat production, which is the dominant source term in human thermoregulatory physics during physical

activity [42, 43]. At the same time, under environmental heat stress, cardiac output is redistributed to the skin to facilitate heat dissipation [44]. Consequently, heart rate reflects both the rate of internal heat generation (metabolic rate) and the circulatory effort required for thermoregulation [45]. Grounded in these mechanisms, variations in heart rate provide a physiological link to human thermoregulation, forming the basis for CBT estimation approaches.

Importantly, heart rate is one of the most accessible and reliable physiological signals available from wearable devices. It can be measured continuously, non-invasively, and with high signal quality across a wide range of activities and environments. These practical advantages, combined with its physiological relationship to human thermoregulation, make it widely adopted as a surrogate input for CBT estimation in wearable-based monitoring systems [46].

To overcome the limitations of direct measurement, non-invasive CBT estimation methods have been developed primarily in two complementary directions. One direction is based on thermophysiological modeling, in which CBT is predicted from first-principles representations of heat production, storage, and transfer within the human body. In their classical formulations, such off-line models do not rely on wearable physiological signals such as heart rate, but instead require physiological parameters, environmental conditions, activity type or metabolic rate, as well as clothing characteristics [47, 48]. These parameters are difficult to measure or determine continuously in free-living conditions, which has historically limited the direct deployment of first-principles models for real-time wearable monitoring.

The other direction is the data-driven estimation frameworks grounded in empirical relationships between wearable physiological signals and CBT to enable real-time tracking. State-space models, statistical/machine learning methods, and hybrid approaches have been explored to balance accuracy, robustness, physiological interpretability, adaptability, and computational efficiency while remaining compatible with the sensing capabilities of commercial wearable devices [38, 46, 49, 50].

These developments illustrate a complementary methodological pathway in which CBT is estimated from wearable physiological signals when direct measurement

is impractical, enabling continuous monitoring of internal thermal strain in sport, occupational, and military applications.

1.2 Research objectives

At present, no universally accepted numerical accuracy criterion has been established specifically for algorithms that estimate CBT from wearable physiological signals during physical activity. The ± 0.3 °C requirement specified in ISO 80601-2-56:2017 concerns the technical accuracy of clinical thermometers when compared with calibrated reference temperature sources, rather than the agreement between an algorithmic estimate and real CBT [51]. It may therefore provide a useful numerical reference for interpreting the magnitude of estimation errors, but should not be applied directly as a formal acceptance threshold for the methods considered in this thesis. Although relatively low estimation errors have been reported for heart-rate based methods in individual studies [46, 52], reported performance varies with the study population and experimental protocols [53]. Consequently, it remains unclear whether current methods provide sufficiently consistent accuracy across different conditions. Further methodological development is therefore needed not only to reduce average estimation error, but also to improve robustness, generalizability, real-time feasibility, and physiological interpretability. Accordingly, although it is not treated as a formal clinical acceptance criterion, an RMSE of approximately 0.3 °C is adopted in this thesis as a pragmatic research target for CBT estimation under controlled physical-activity conditions.

Motivated by the limitations of direct temperature measurement, specialized sensing hardware, and first-principles models, this thesis examines how the internal thermal state can be reliably inferred from wearable physiological signals. The overall aim of this thesis is to develop non-invasive CBT estimation methods using limited wearable physiological data that progressively reduce estimation error towards, and where possible below, an RMSE of approximately 0.3 °C under controlled physical activity conditions, while also improving robustness and physiological interpretability, and maintaining practical computational feasibility.

Specifically, this thesis addresses the following research objectives:

- To evaluate the accuracy of existing CBT estimation methods using wearable physiological signals
- To investigate modeling strategies that capture the dynamic relationships between wearable physiological measurements and CBT across different activity phases, with the aim of reducing RMSE towards approximately 0.3 °C.
- To examine how to balance the physiological interpretability, accuracy, robustness and computational efficiency of the CBT estimation frameworks.

1.3 Contributions and organization of the thesis

The core chapters (Chapters 2–7) are based on peer-reviewed journal publications. Unless otherwise stated, all computational implementation, algorithm development, model training, testing, statistical analyses, and generation of results presented in this thesis were undertaken by the candidate. This included the re-implementation of the benchmark models evaluated in Chapter 3 and the development of the computational frameworks presented in Chapters 4–7. Contributions from collaborators are described in the relevant chapters.

All computational analyses presented in Chapters 3–7 were based on two source datasets, referred to throughout this thesis as Dataset 1 and Dataset 2. Different chapters used different subsets of participants, experimental conditions, and input variables according to their methodological requirements, but no additional datasets were introduced. In both datasets, heart rate values were obtained directly from the wearable chest ECG measurement devices rather than calculated from raw ECG signals in this thesis. The availability and access conditions of the two datasets have been clarified in Appendix A.

The organization of the thesis and the associated research outputs are detailed below.

Chapter 2 presents a systematic review of non-contact body temperature measurement methods, including infrared thermometers and thermal scanners. Conventional reference methods for CBT measurement, such as rectal and gastrointestinal thermometry, can provide reliable measurements but are invasive and impractical for

continuous monitoring during physical activity. Chapter 2 therefore examines whether non-contact infrared techniques could provide a more practical alternative. The review shows that, although these techniques have potential in static screening and other controlled settings, their accuracy and reliability are insufficient for continuous CBT monitoring under dynamic conditions. These limitations motivate the subsequent focus on indirect CBT estimation from wearable physiological signals. This chapter is based on the following published work:

Y. Zhao and J. H. Bergmann, “Non-contact infrared thermometers and thermal scanners for human body temperature monitoring: A systematic review,” *Sensors*, vol. 23, no. 17, Art. no. 7439, 2023.

Chapter 3 provides a systematic review of data-driven CBT estimation methods during physical activities. It outlines the principal modeling paradigms reported in the literature, including statistical models, Kalman filter and extended Kalman filter (KF/EKF) approaches, machine learning models, and hybrid methods combined physics-based models with data-driven components, and summarizes their methodological details, input requirements, and reported performance. In addition, this chapter conducts a comprehensive benchmark evaluation of the selected methods using two independent datasets, enabling a consistent and quantitative comparison of their performance. This chapter establishes a clear methodological basis for the subsequent improvements developed in this thesis. This chapter is based on the following published work:

Y. Zhao, W. Li, and J. H. M. Bergmann, “Data-driven estimation of core body temperature during physical activity under heat stress: A systematic review and standardized evaluation,” *Building and Environment*, vol. 297, Art. no. 114591, 2026.

Chapters 4–7 develop a progressive methodological pipeline for CBT estimation during physical activities under heat stress. Based on the methodological basis established in Chapter 3, these chapters advance from classical extended Kalman filter to probabilistic state modeling, then to the adaptive correction, and finally to physics-guided hybrid learning.

Chapter 4 extends classical EKF-based CBT estimation by introducing a multi-model extended Kalman filtering framework with variance-based fusion using heart

rate only, including two approaches: label-informed (PSSK) method and unsupervised trial-clustering (TCBK) method. Their performance is evaluated within and across datasets. This chapter is based on the following published work:

Y. Zhao and J. H. M. Bergmann, “Core body temperature estimation from heart rate via multi-model Kalman filtering and variance-based fusion,” *Physiological Measurement*, vol. 46, no. 10, Art. no. 105002, 2025.

Chapter 5 moves from heuristic variance-based fusion and fixed-label model to probabilistic state inference by coupling a Hidden Markov model (HMM) with a particle filter (PF). The chapter also adopts the online adaptation and improved genetic resampling to improve accuracy using only heart rate. This chapter is based on the following published work:

Y. Zhao and J. H. Bergmann, “Real-time estimation of core body temperature for heat stress monitoring in hot environments using wearable heart rate sensors,” *Building and Environment*, vol. 293, Art. no. 114368, 2026.

Chapter 6 presents a residual-compensated adaptive EKF (RCAKF) framework in which observation model errors are compensated using a Long Short-Term Memory (LSTM) network. By combining residual compensation with adaptive noise estimation, the proposed framework improves the overall accuracy of CBT estimation. This chapter is based on the following published work:

Y. Zhao and J. H. M. Bergmann, “Estimating core body temperature from heart rate using a residual-compensated adaptive Kalman filter,” *Biocybernetics and Biomedical Engineering*, vol. 45, pp. 617–629, 2025.

Chapter 7 develops a physics-guided hybrid framework that integrates a two-node thermoregulation model with neural network residual compensation for CBT estimation. This chapter systematically compares the hybrid variants with physics-only, EKF-based, and purely data-driven baselines to evaluate the accuracy, robustness, and model complexity for wearable applications. This chapter is based on the following published work:

Y. Zhao and J. H. Bergmann, “A physics-guided hybrid model for core body temperature estimation: Two-node thermoregulation with residual learning,” *Expert Systems with Applications*, vol. 329, Art. no. 133001, 2026.

In addition to the studies incorporated in Chapters 2–7, the candidate has contributed to several related peer-reviewed publications during the doctoral period that are not included in this thesis:

Y. Zhao and J. H. M. Bergmann, “Computational and experimental study of thermal behaviour in the oral cavity for biosensing applications,” *Results in Engineering*, vol. 26, Art. no. 105284, 2025.

Y. Zhao and J. H. M. Bergmann, “A hybrid core temperature estimation method integrating physiological and environmental factors,” *IEEE Sensors Letters*, vol. 9, no. 7, Art. no. 7003604, 2025.

Y. Zhao, L. de Almeida e Bueno, D. A. Holdsworth, and J. H. M. Bergmann, “Evaluating the agreement between oral, armpit, and ear temperature readings during physical activities in an outdoor setting,” *International Journal of Environmental Research and Public Health*, vol. 21, no. 5, Art. no. 595, 2024.

2

Non-Contact Infrared Thermometers and Thermal Scanners for Human Body Temperature Monitoring: A Systematic Review

As outlined in Chapter 1, conventional reference methods for CBT measurement, such as rectal and gastrointestinal thermometry, are invasive and impractical for continuous monitoring during physical activity. This motivates the investigation of less burdensome alternatives. Before investigating the indirect estimation models in subsequent chapters, it is essential to first evaluate the performance of direct non-contact sensing technologies widely used for screening.

This chapter presents a systematic review of non-contact infrared thermometers (NCITs) and infrared thermography (IRT), synthesising evidence from 72 unique settings across 32 studies ranging from climate-controlled investigations to clinical applications. The findings indicate that while NCITs and IRT displayed positive correlations with contact-based reference measurements, their accuracy—as assessed relative to these reference standards—is highly conditional. NCITs demonstrated slightly superior accuracy over IRT: 29 of 50 NCIT settings and 4 of 22 IRT settings reported agreement within the commonly adopted threshold of $\pm 0.3^{\circ}\text{C}$, which is used as a descriptive reference based on technical accuracy requirements against calibrated blackbody sources.

The performance was found to be influenced by measurement location, measurement distance, sensor type, environmental conditions and application scenarios. It is also important to note that the contact-based reference methods used across

the reviewed studies themselves exhibit variability and uncertainty during physical activities, which further constrains the achievable agreement of non-contact techniques. Taken together, these results suggest that while non-contact infrared sensing can be useful for static screening, it lacks the robustness required for continuous monitoring under dynamic field conditions. This limitation underscores the necessity of the contact-based methods.

2.1 Introduction

Body temperature is a vital sign that reflects a person’s health status. The regulation of body temperature is controlled by the hypothalamus, which uses the nervous system and body fluid to control heat production and heat dissipation in order to maintain the dynamic balance of body temperature [54, 55]. Maintaining the appropriate body temperature level is crucial for the human body and normally ranges between 36.16 °C and 37.02 °C for healthy individuals [56]. Changes in body temperature can indicate the occurrence of different diseases. This change is a response to stress on the nervous and immune systems to pathogens, such as bacteria or viruses [57].

The importance of body temperature monitoring has been recognized since ancient times, with physicians in the early Roman Empire noting a relationship between body temperature and poor health [58]. The history of temperature measurement dates back centuries, with the development of various devices and techniques to monitor body temperature. Galileo Galilei invented one of the earliest thermometers in the late 16th century, and the mercury-in-glass thermometer was developed by Gabriel Fahrenheit in the early 18th century [59, 60]. These early devices laid the foundation for modern temperature measurement techniques.

However, traditional contact thermometers, such as mercury or electronic thermometers, have limitations that can impact the accuracy and safety of temperature monitoring. For example, contact thermometers can affect the temperature field of the object itself when in touch with the object, resulting in potential temperature measurement errors [61, 62]. In addition, contact methods require physical contact

with the person being monitored, which can increase the risk of cross-infection and pose practical challenges for large-scale monitoring efforts [63, 64].

The development of infrared technology since the mid-20th century introduced new possibilities for non-contact temperature sensing. This technology relies on the detection of infrared radiation emitted by objects to determine their temperature. Non-contact infrared thermometers (NCITs) and thermal scanners use this principle to measure the temperature of the skin surface without coming into contact with the skin. This method of temperature measurement can provide an accurate reading within seconds and it can be performed at a distance from several centimeters to meters away, reducing the risk of infection transmission [63]. Non-contact thermal screening, including the application of infrared thermography (IRT) and non-contact infrared thermometers (NCITs), is deemed a safe tool for temperature screening in the outbreak of infectious diseases [65].

Since the outbreak of COVID-19, non-contact temperature screening methods have been widely used in crowded places such as airports, markets, stations, and hospitals [66]. The pandemic has highlighted the need for rapid and accurate fever detection to mitigate the spread of the virus. This has led to an increased demand for non-contact temperature measurement devices and a surge in research evaluating their performance. While non-contact infrared thermometers and thermal scanners have gained widespread popularity in recent years, their accuracy and reliability for detecting fever in humans are still a subject of debate. Several studies have suggested that non-contact tools are a reliable method with convenience and efficiency [63, 64, 66]. However, others have suggested that the accuracy and robustness of this method is not satisfying clinically and that contact sensors remain the preferred method of measurement [67, 68].

To address this issue, a systematic review of the relevant literature was conducted to evaluate the quantitative accuracy and reliability of non-contact infrared thermometers and thermal scanners compared to reference standard tools, such as electronic thermometers, which often consist of contact-based methods. This review aims to synthesize the available evidence on the performance of these devices and to provide a critical appraisal of the studies that have investigated their use for

body temperature measurement and fever screening. By conducting this systematic review, it is hoped to provide guidance for appropriate sensor selection and future research in this area, as well as inform policy decisions related to the use of either contact or non-contact body temperature measurement methods.

2.2 Materials and Methods

2.2.1 Search Strategy

The structure of the content below follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [69]. A systematic literature search was conducted in multiple databases, including the Web of Science Core Collection, PubMed, and IEEE Xplore, to identify studies published before 31 March 2023 that reported the use of non-contact infrared body temperature measurement methods. The following keywords and search terms were used:

(Core Temperature OR Body Temperature) AND (measure* OR predi* OR monit* OR estimat*) AND (remote* OR camera* OR wireless* OR non-contact* OR non contact*)

The keywords were selected to cover three main elements: body temperature, measurement and monitoring, and non-contact sensing. Broad terms and truncation were used to maximise the retrieval of potentially relevant studies, with final relevance determined during study screening.

The search was conducted by Zhao and Bergmann, who screened the titles and abstracts of the identified studies to assess their relevance. Data were extracted from the eligible studies using a standardized form. The quality of the included studies was assessed using the Cochrane Risk of Bias tool [70].

2.2.2 Study Selection Criteria

Studies that met the following criteria were included:

1. Comparison of non-contact infrared body temperature measurement with an established reliable measurement reference for body temperature (e.g. contact measurements of axillary, tympanic, and oral temperature);

2. Study conducted on human participants of any age, gender, or ethnicity;
3. Reporting quantitative measurements of body temperature;
4. Published in peer-reviewed journals;
5. Written in English.

Papers that did not meet these criteria were excluded. For this review, non-contact temperature measurement devices were defined as infrared thermometers, infrared thermal scanners, and infrared thermal cameras. The reference temperature was defined as the temperature obtained by other body temperature measurement sensors (which might also be a non-contact tool itself).

2.2.3 Data Extraction

Two independent reviewers extracted data from each included study using a standardized data extraction form. The following data were extracted:

1. **Basic information of the study:** author, year of publication.
2. **Study methodology:** place of experiment, type and brand of non-contact device used, environment temperature, test distance and measurement location.
3. **Population characteristics:** group size, age structure, gender ratio.
4. **Reference temperature:** type and brand of reference tool used, measurement location of reference temperature.
5. **Quantitative results:** measured temperature range, correlation coefficient between temperature obtained by non-contact devices and reference temperature, mean difference (MD) of results between non-contact tools and reference tools, and any other reported metrics of accuracy or reliability.

The mean difference (MD) is defined as [51]:

$$MD = \frac{\sum_{i=1}^n (T_{\text{dev},i} - T_{\text{ref},i})}{n} \quad (2.2.3.1)$$

where i is the index number of a single measurement; n is the total number of the measurements; $T_{\text{dev},i}$ is the recorded temperature using the non-contact device from the i th measurement; $T_{\text{ref},i}$ is the recorded temperature using the specific reference tool from the i th measurement.

The standard deviation (SD) of the MD is defined as:

$$SD = \sqrt{\frac{\sum_{i=1}^n [(T_{\text{dev},i} - T_{\text{ref},i}) - MD]^2}{n - 1}} \quad (2.2.3.2)$$

where i is the index number of a single measurement; n is the total number of the measurements; $T_{\text{dev},i}$ is the recorded temperature using the non-contact device from the i th measurement; $T_{\text{ref},i}$ is the recorded temperature using the specific reference tool from the i th measurement; MD is the mean difference between non-contact tools and reference tools as calculated in Equation (2.2.3.1).

The full standardized data extraction forms are provided in Supplementary Tables S1 and S2.

Included studies might contain experiments with more than one non-contact device or reference tool. Each reported pair of non-contact device and reference tool was regarded as a separate test and thus one study might have multiple entries within the generated table.

The mean difference (bias) and the standard deviation (SD) between non-contact and reference systems were considered to be the primary outcome for the qualitative assessment of the included studies. If a study reported the mean absolute error (MAE) or root mean square error (RMSE) instead of the mean difference, it was not included in the quantitative analysis. The supplementary table also includes any alternative performance measures that are applied in, e.g., fever screening, such as the sensitivity, specificity, area under the curve (AUC), positive predictive value (PPV), and negative predictive value (NPV).

2.2.4 Quality Assessment

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool was used to assess the quality of the included studies [71]. The quality assessment

was performed by two independent reviewers, and any discrepancies were resolved through discussion.

2.2.5 Data Synthesis

A narrative synthesis of the available evidence was conducted due to the anticipated heterogeneity of the included studies. The characteristics and findings of each study were summarized, and any patterns or inconsistencies in the results were identified. By following these methods, it is aimed to provide a comprehensive and transparent review of the available evidence on the comparison of non-contact infrared body temperature measurement with contact body temperature sensing.

2.3 Results

A total of 6087 articles were identified through the database search, of which 1662 were duplicates. After screening the titles and abstracts of the remaining 4295 articles, 130 records were selected for full-text screening. However, not all articles were obtainable from the University library system, online resources, or by contacting corresponding authors. A total of 122 records out of these 130 records were assessed based on their full text, as no access could be obtained for 8 identified manuscripts. Finally, 58 studies were included in the systematic review and 32 studies were included in the quantitative analysis (see Figure 2.1), as 26 studies had no reported MD or SD.

2.3.1 Characteristics of Included Studies

A total of 32 studies were included in the quantitative analysis after screening 4295 records and assessing 122 full-text articles. The studies were conducted in various scenarios, including hospitals, rooms, and laboratory settings. These 32 studies contain 72 different experiment settings. The sample sizes ranged from a single person to 1,113 participants, and the studies were conducted in different countries around the world. A total of 20 studies reported the range of their environment temperature, which varied from 9 °C to 38.8 °C. The included studies contain two non-contact body temperature devices: the non-contact infrared thermometers (NCITs) and infrared scanners, which use infrared thermography (IRT). A total of 22

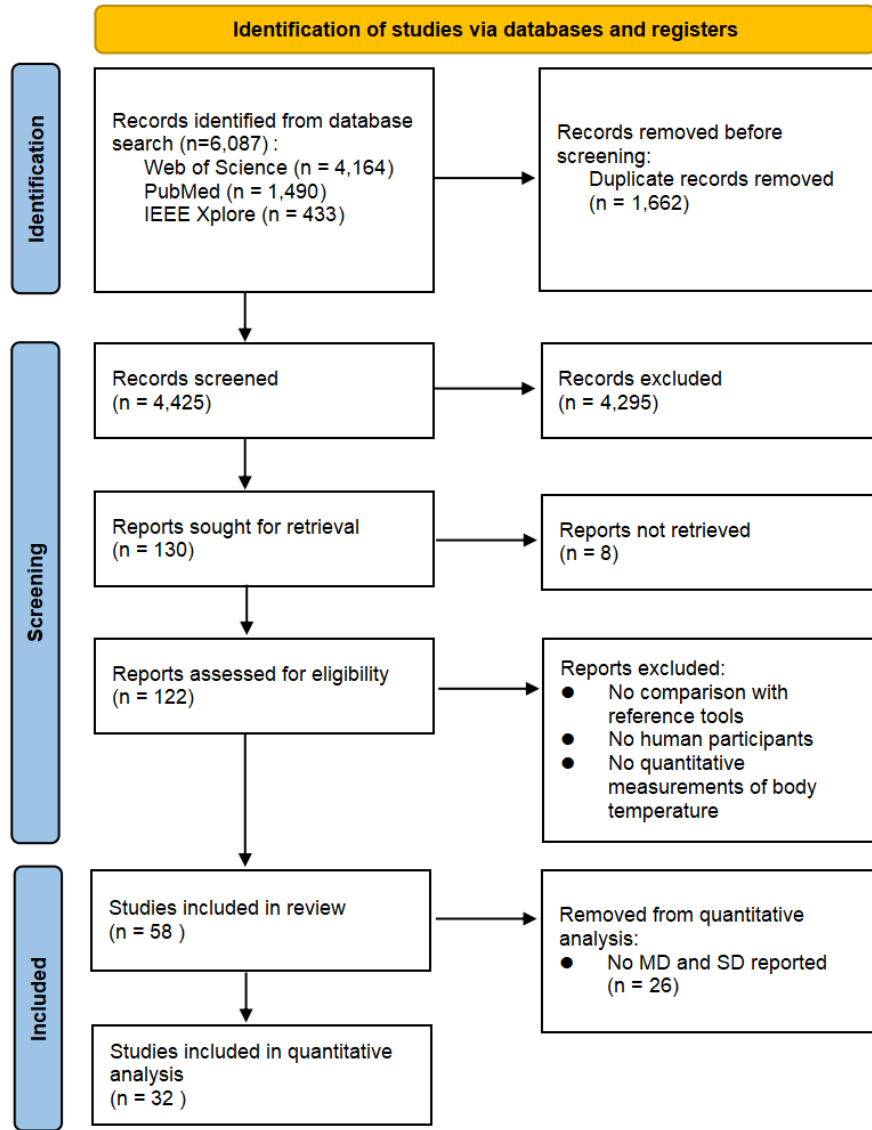


Figure 2.1: Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) flow diagram of the study selection process.

studies (50 settings) used NCIT, and 13 studies (conducted in 22 settings) used IRT. A total of 28 studies reported the number of participants by gender. Among the 6792 participants of these 28 studies, 3426 (50.4%) were men and 3366 (49.6%) were women.

Sensor Types

Figure 2.2 presents a forest plot summarising the mean difference (MD) $\pm$ standard deviation (SD) reported across 22 studies (50 experimental settings) using non-contact infrared thermometers (NCITs) for body temperature measurement. The measurement body sites varied across studies, ranging from the forehead to

peripheral locations such as the fingers. Overall, the results demonstrate substantial variability in the accuracy and reliability of NCITs when compared with contact-based reference methods.

The highest absolute MD was 2.34 °C (SD = 1.06 °C), observed in measurements taken from the forehead of children aged between 1 month and 18 years, with tympanic thermometry used as the reference. In contrast, the lowest absolute MD was 0.02 °C (SD = 0.388 °C), reported for forehead measurements in healthy newborns, again compared against a tympanic thermometer. Across all experimental settings, 21 out of 50 settings exhibited a positive MD, while the remaining 29 settings showed a negative MD. When the setting-level results were pooled using the reported participant numbers, the overall pooled result was -0.181 ± 0.779 °C, indicating a small average underestimation but substantial variability across settings.

Figure 2.3 shows the forest plot of MD $\pm$ SD across 13 studies (22 experimental settings) that employed infrared thermography (IRT) for body temperature assessment. The reported measurement sites primarily included facial and general skin regions, with a wide range of reference thermometry methods.

The largest absolute MD was 11.2 °C (SD = 2.056 °C), recorded from skin temperature measurements in a cohort of 61 adults (mean age $\pm$ SD: 67 ± 15.2 years), using an oral probe as the reference method. Conversely, the smallest absolute MD was 0 °C (SD = 0.2 °C), reported for facial measurements in a paediatric cohort comprising 154 children with COVID-19 and 147 healthy controls, with axillary mercury thermometry as the reference. Across all IRT-based studies, 9 settings yielded a positive MD, 1 setting reported an MD of zero, and the remaining 12 settings exhibited a negative MD. The overall pooled result was -1.601 ± 3.797 °C, indicating a larger average underestimation and substantially greater variability than observed for NCITs.

Reference Temperature

The reference temperature utilized for comparison in the included studies varied, with measurements taken at different body locations, such as tympanic, axillary, oral, skin, rectal, and gastrointestinal sites. The selection of the reference temperature

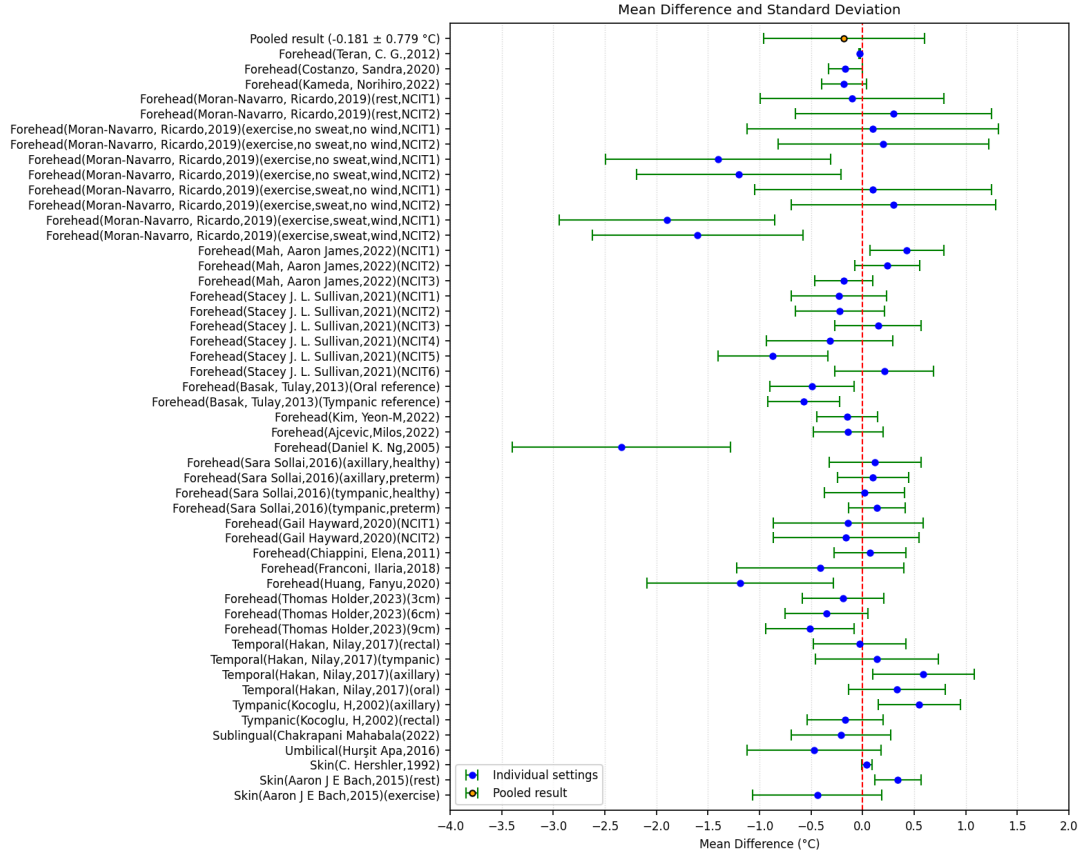


Figure 2.2: Forest plot comparing MD $\pm$ standard deviation of included studies of NCITs [72–93].

could have influenced the evaluation of the accuracy and reliability of the non-contact infrared temperature measurement methods.

Among the included studies, 11 studies (comprising 14 settings) used axillary temperature as the reference body site, employing both digital or mercury thermometers (Figure 2.4). Out of these 14 settings, 6 had an MD > 0 °C, 1 had an MD = 0 °C, and the remaining 7 had an MD < 0 °C. The IRT study conducted by Unursaikhan Batbayar demonstrated the best MD and SD [96]. This study was conducted at the Central Hospital of Songinokhairkhan District in Mongolia; although the environmental temperature was not reported, the experiment was carried out under stable clinical conditions. The study included a group size larger than 100, with both male and female participants. The region of interest for body temperature was extracted using camera image processing algorithms. Conversely, the two settings of the IRT study conducted by Peter Y. Chan exhibited the worst MD and SD [103]. In this study, the infrared measurements were reported without

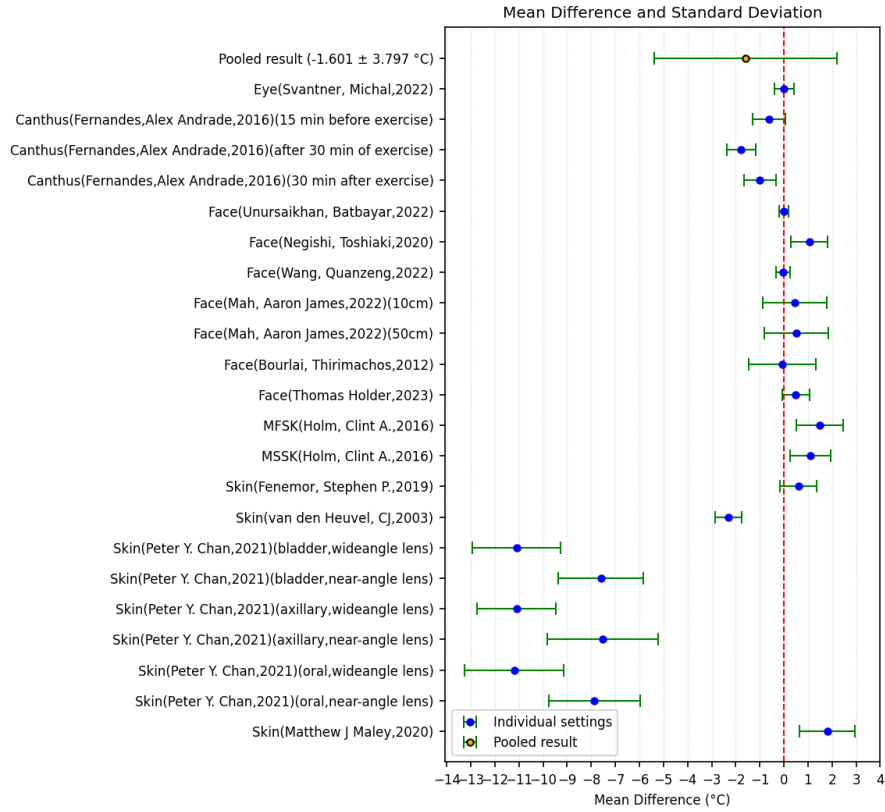


Figure 2.3: Forest plot comparing MD $\pm$ standard deviation of included studies of IRT [76, 87, 94–104].

the software-based compensation commonly used to correct the systematic offset of non-invasive temperature measurements, which likely contributed to the large MD greater than 7 °C.

In seven studies (comprising eight settings), the reference temperature was measured using tympanic thermometers (Figure 2.5). Among these eight settings, three had an MD > 0 °C and five had an MD < 0 °C. One setting of the NCIT study conducted by Sara Sollai had the best MD and SD. This particular setting involved 70 pre-term newborns (28 boys and 42 girls) and was carried out in incubators [82]. On the other hand, the NCIT study conducted by Daniel K. Ng had the worst MD and SD [81]. This study was conducted at Kwong Wah Hospital and included 567 patients aged between 1 month and 18 years old.

Regarding oral reference temperature measurements, 7 studies (covering 16 settings) used the oral cavity, employing various tools such as chemical dot thermometers, digital thermometers, mercury thermometers, or oral probes (Figure 2.6). Among these 16 settings, 7 had an MD > 0 °C and 9 had an MD < 0 °C. One setting of

the IRT study conducted by Wang Quanzeng had the best MD and SD [98]. This study involved 1020 participants and was conducted at the Health Center of the University of Maryland. The overall ambient temperature ranged from 20 to 29 °C. The region of interest for body temperature in the IRT images was identified by matching landmarks on visible light images to thermal images using an image registration approach and manual labeling. Conversely, the two settings of the IRT study conducted by Peter Y. Chan exhibited the worst MD and SD [103]. These settings followed similar experimental conditions as the aforementioned axillary temperature findings reported in the same study.

For skin temperature (non-axillary sensor location), 8 studies (comprising 11 settings) used this body site as the reference, employing tools such as zero-heat-flux cutaneous thermometers, thermistors, skin temperature probes, digital thermometers, and thermocouples (Figure 2.7). Among these 11 settings, 5 had an MD > 0 °C and 6 had an MD < 0 °C. One setting of the NCIT study conducted by C. Hershler had the best MD and SD, involving five male adults. Conversely, the two settings of the IRT study conducted by Matthew J. Maley exhibited the worst MD and SD. This study involved 52 male adults and was conducted in a climate-controlled chamber with an environmental temperature of 30.3 ± 0.9 °C.

Additionally, three studies reported rectal temperature as the reference, using mercury thermometers and rectal probes. Four studies used gastrointestinal temperature, measured by ingestible sensors, as the reference temperature. One study utilized bladder temperature measured via the bladder catheter. The choice of reference temperature, as well as the calibration and maintenance of non-contact temperature measurement devices, are crucial factors to consider when assessing the accuracy and reliability of NCIT and IRT devices. These factors can directly impact the performance of non-contact temperature measurement methods in various clinical and non-clinical settings (Figures 2.4–2.7).

Calibration of Non-Contact Devices

The reporting of device calibration varied across the included studies. Among the 22 studies that evaluated NCITs, 7 reported that the sensing device had been calibrated or verified. Of these, 3 described the calibration procedure, primarily

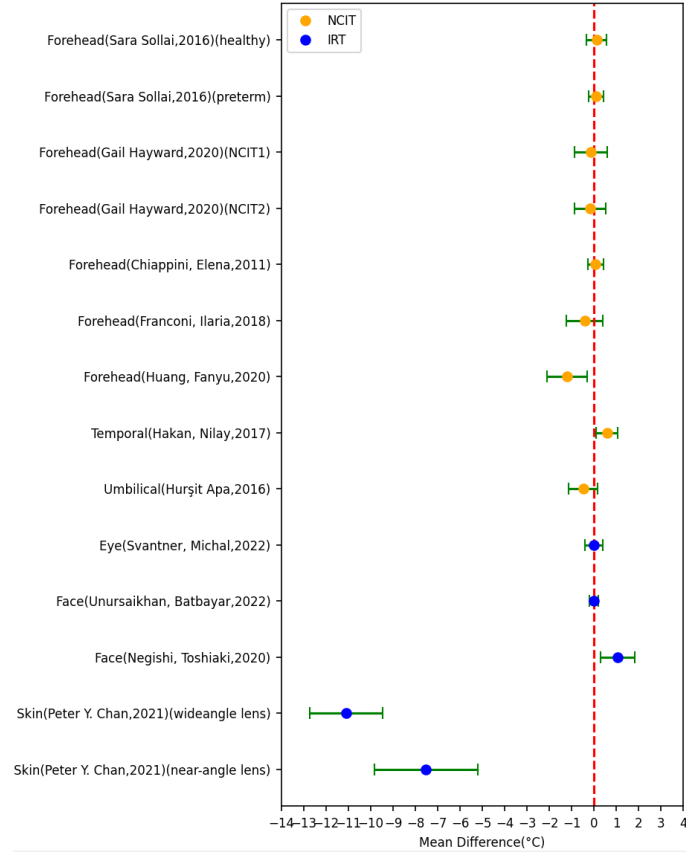


Figure 2.4: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with axillary temperature as reference [83–88, 91, 94, 96, 97, 103].

involving calibration against a blackbody reference source, whereas 4 mentioned calibration without providing sufficient methodological details. The remaining 15 studies did not report whether device calibration had been performed. Among the 13 studies that evaluated IRT systems, 5 reported calibration of the thermal imaging equipment, while 8 did not report any calibration procedure. Where calibration methods were described, blackbody reference sources were most commonly used. As some studies evaluated both NCIT and IRT devices, these counts are reported separately by device category rather than as mutually exclusive study totals.

Environment Temperature, Average Body Temperature, and Average Age

Among all of the 72 settings from the included 32 studies, 43 reported an average environmental temperature ranging from 20.1 °C to 38.8 °C. Figure 2.8 shows the mean difference between non-contact measurements and reference data against environmental temperature. A total of 38 settings reported the average

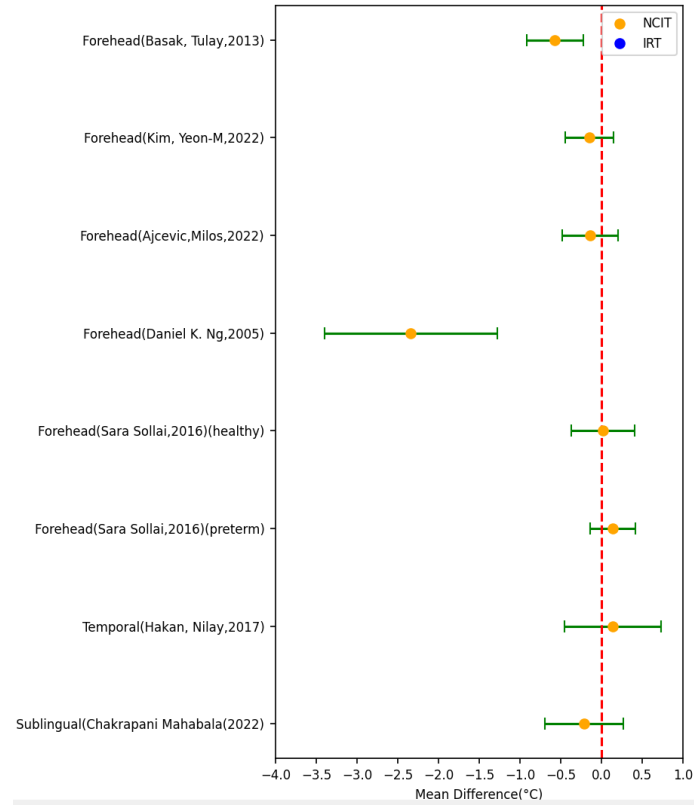


Figure 2.5: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with tympanic temperature as reference [78–82, 88, 90].

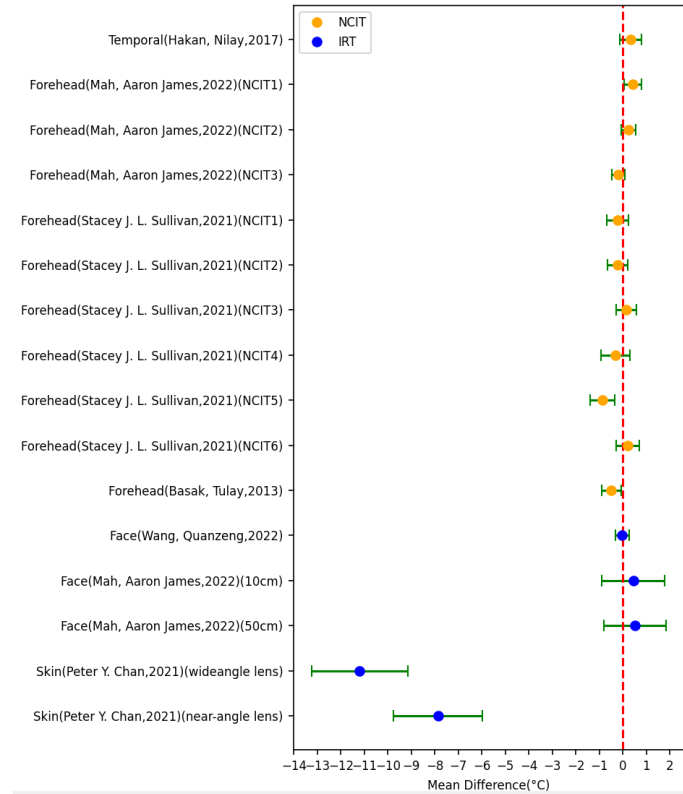


Figure 2.6: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with oral temperature as reference [76–78, 88, 98, 103].

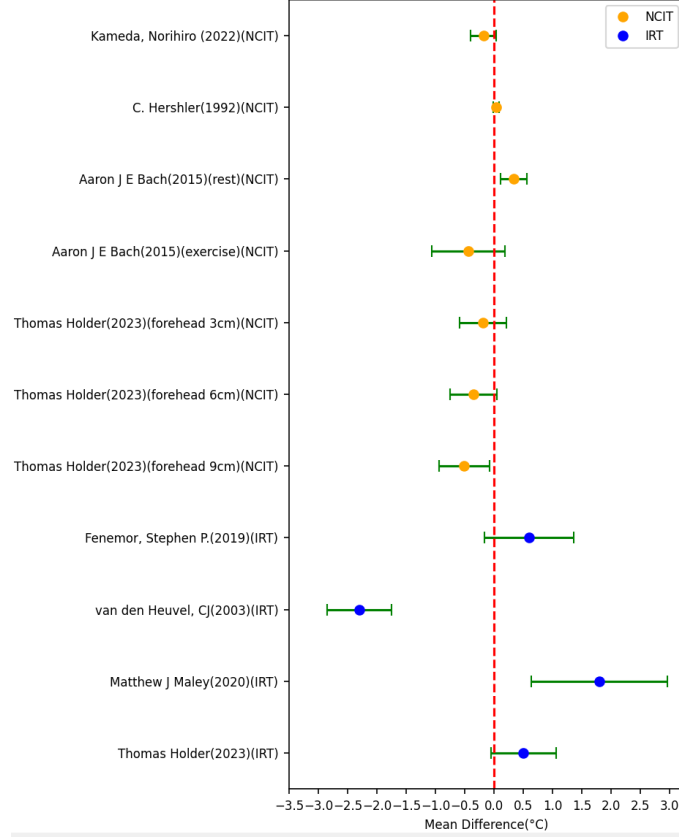


Figure 2.7: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with skin temperature as reference [74, 87, 92, 93, 101, 102, 104].

reference temperature varied from 34.2 °C to 38.3 °C. Figure 2.9 shows the mean difference between non-contact and reference data when plotted against the average body temperature as obtained by the reference sensor. The lowest average body temperature (34.2 ± 1.0 °C) was reported by Fenemor, Stephen P [101], and Moran-Navarro, Ricardo reported the highest average body temperature, with a value of 38.3 ± 0.9 °C [75]. A total of 24 settings reported the average age of the subjects varied from 14.6 months to 67 years. Figure 2.10 depicts the mean difference between the non-contact devices and reference tools in relation to the subjects' average age.

2.3.2 Methodological Quality of Included Studies

The Quality Assessment of Diagnostic Accuracy Studies (QUADAS-2) tool was used to assess the quality of all the included studies [105]. The risk of bias was assessed across four key domains, including patient selection, index test, reference standard, as well as flow and timing. Meanwhile, the patient selection, index test, and reference

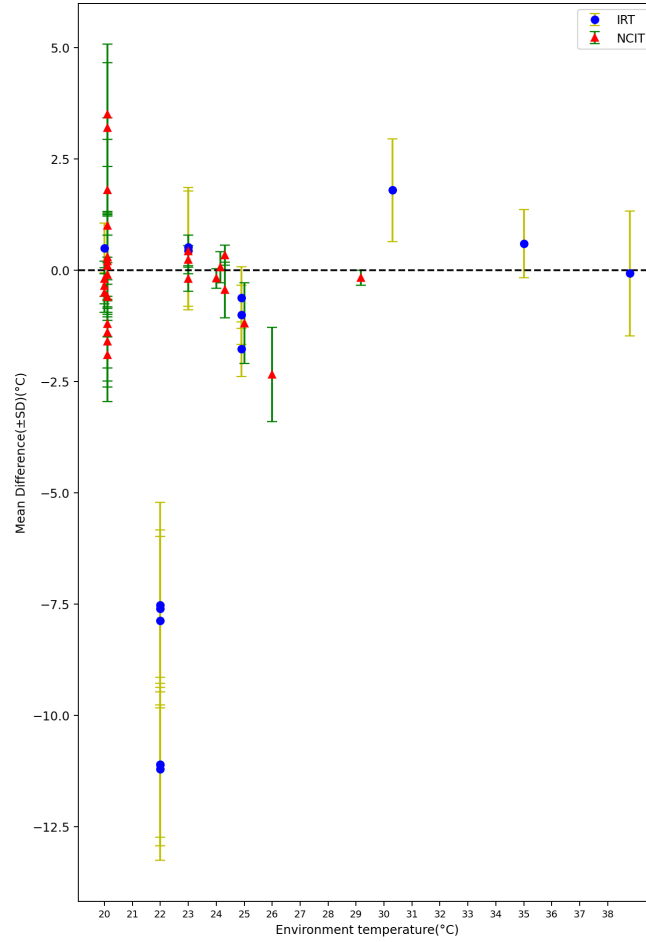


Figure 2.8: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with different environment temperature.

standard were also assessed for concerns regarding applicability. Supplementary Tables S1 and S2 summarize the results of the quality of the included studies [106].

Across the included studies, the risk of bias was generally low or unclear for patient selection and the reference standard, but was often high or unclear for the index test. This was mainly because blinding between the non-contact and reference measurements was not reported, and fever-detection thresholds were not always clearly pre-specified, which could lead to overestimation of diagnostic performance. Applicability concerns were also frequently high for the reference standard because the studies used different measurement sites and devices, which do not necessarily represent the same physiological temperature and therefore limit direct comparison across studies.

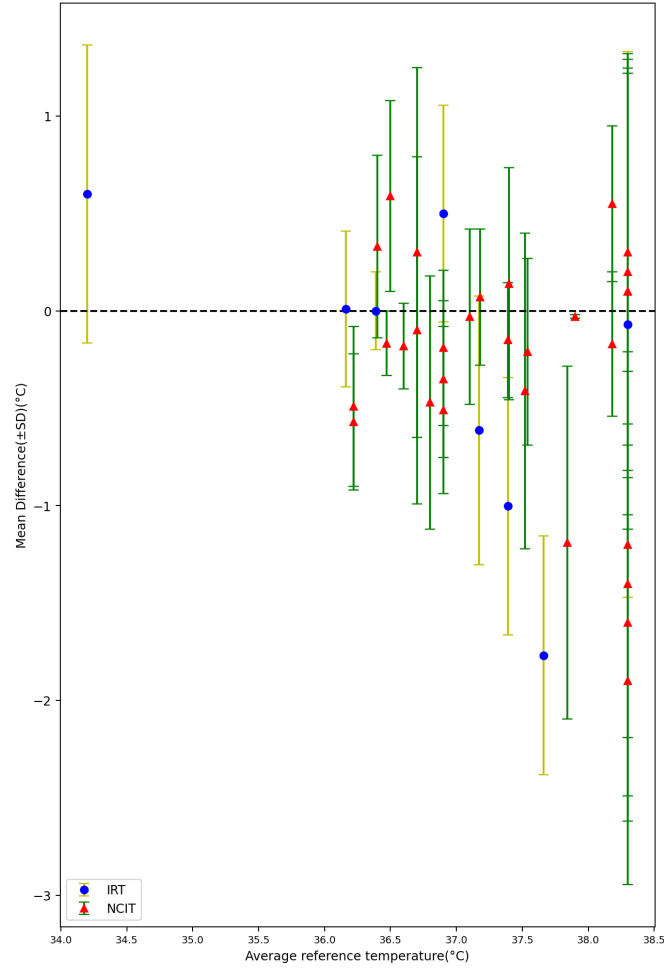


Figure 2.9: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with different average body temperature measured by reference tools.

2.4 Discussion

Non-contact infrared thermometers (NCITs) and infrared thermography (IRT) have been widely recognized for their efficiency and safety in detecting body temperature, yielding considerable benefits in healthcare settings. For example, by employing NCITs, nursing efficiency can be improved while minimizing patient discomfort and distress [107]. This review provides an overview of the performance of these devices. This systematic review aimed to evaluate the accuracy and reliability of NCITs and IRT for human body temperature monitoring in various settings and conditions. A total of 32 studies were included, which demonstrated varying levels of accuracy and reliability for both NCIT and IRT devices when compared to reference standard tools, which were often based on contact sensing.

The results show that the accuracy of NCITs (pooled MD= -0.181 ± 0.779 °C)

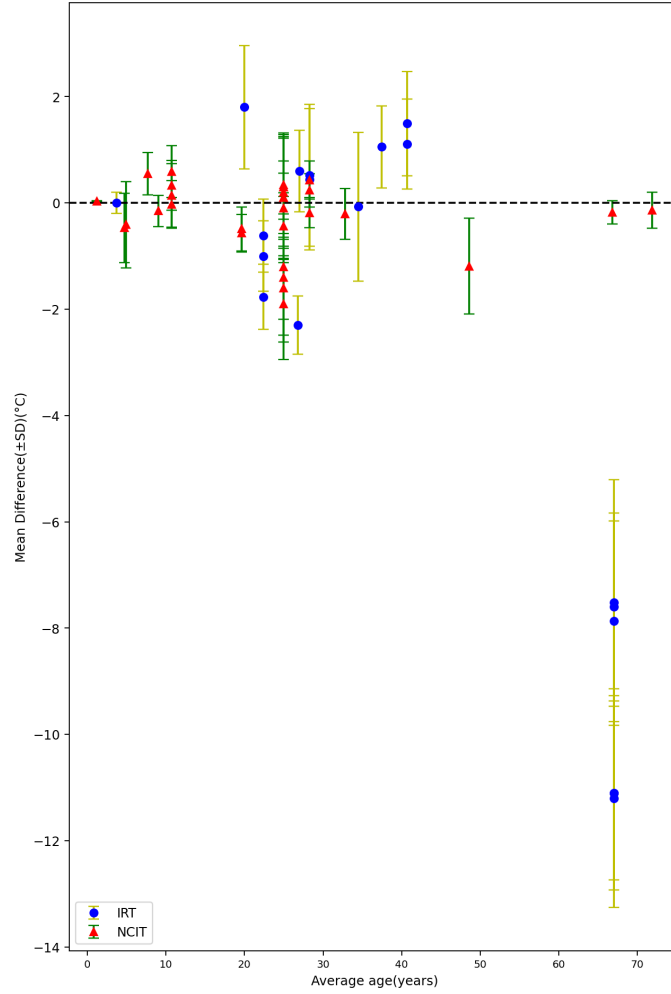


Figure 2.10: Forest plot comparing mean difference $\pm$ standard deviation (SD) of included studies with different average age of subjects measured by reference tools .

is better than that of IRT (pooled MD= -1.601 ± 3.797 °C). Generally, NCITs have a shorter effective distance and are less affected by external factors than IRT. Although, NCITs often exhibit higher accuracy and reliability compared to IRT, IRT still has its own unique advantages. IRT can be used to quickly screen the body temperature of large groups of people from longer distance. For example, Armote Somboonkaew’s research shows that IRT can screen the body temperature of nine people simultaneously at a speed of eight frames/second [108] and a correlation coefficient of 0.731 was found between IRT and forehead skin thermometers. It achieves 100% sensitivity and 92.6% specificity under the cutoff value of 37.5 °C. Sun deployed thermal imagers at airports for body temperature screening, testing 617,289 individuals and identifying 391 individuals with fever, with a mean difference of 2.7 °C lower than the axillary temperature [109].

The comparison results of different reference positions show that the measurement results of NCITs and IRT are not consistently higher or lower than those of a specific reference position. However, the wide range of mean differences indicate substantial variability in the performance of these devices. Factors such as the measurement body site, sensor type, reference standard, and individual differences may have contributed to these discrepancies.

The ISO 80601-2-56:2017 standard outlines the laboratory accuracy requirements for clinical thermometers [51]. Specifically, this accuracy is quantified as the measurement error, which is expressed as the mean difference in readings from a specific temperature source (such as a blackbody radiator or a fluid bath) as measured by both the clinical thermometer and the reference thermometer. For normal use, this measurement error should not exceed ± 0.3 °C. However, the results from clinical studies reveal that this value is sometimes challenging to meet in practice with non-contact sensors. Among all of the 50 NCIT settings and the 22 IRT settings from the 32 articles included in the study, 29 NCIT settings and only 4 IRT settings managed to reach a mean difference within the range of ± 0.3 °C.

The measurement accuracy of NCITs and IRT may be affected by various factors. First of all, the human body's surface is not isothermal, meaning that temperatures can vary across different locations. If the non-contact device measures a different location than the reference device, discrepancies in temperature readings can occur due to the non-isothermal nature of the body's surface. These variations can be influenced by factors such as blood flow, skin thickness, and underlying tissues at a particular measurement site [110]. Secondly, environmental factors, including temperature, humidity, wind speed, and nearby heat sources, can also impact the temperature of the body's surface, thereby affecting measurement accuracy. Normally, the far-infrared emissivity of the body skin is 0.98. In the infrared range relevant to thermal imaging (approximately 8–14 μm), skin emissivity decreases with increasing humidity [68]. The study of Moran-Navarro, Ricardo revealed that the MD between forehead temperature measured using an NCIT and gastrointestinal temperature could reach -1.9 °C in a windy environment, while the MD of the same NCIT was only 0.1 °C without wind [75]. Thirdly, the operation distance of the NCIT and IRT

can also influence the accuracy of body temperature. The accuracy tends to diminish as the distance increases [76, 87]. Additionally, certain characteristics of the human body’s surface, such as skin color, perspiration, hair, and clothing, may influence the body’s heat radiation. These factors can further complicate the measurement process, creating variability in readings and making it more challenging to obtain consistent accurate outcomes [111, 112]. Finally, differences between equipment makes and models can impact the accuracy and consistency of the outputs.

Calibration is a critical determinant of measurement accuracy in infrared sensing. Inadequate or infrequent calibration may introduce systematic bias, reduce measurement consistency, and make temperature estimates sensitive to sensor drift and changes in environmental conditions. Calibration against a reference source, particularly a blackbody source, is therefore important for verifying device performance and ensuring that measurements are comparable across studies [67, 113]. However, calibration was reported in only 7 of the 22 studies evaluating NCITs and 5 of the 13 studies evaluating IRT systems. The limited and incomplete reporting of calibration makes it difficult to determine whether the variability in accuracy across studies reflects differences in the sensing technologies themselves, calibration status, or other experimental factors. Although the absence of reporting does not necessarily mean that calibration was not performed, the relatively small number of studies providing sufficient calibration information represents an important methodological limitation. Future studies should report the calibration method, reference source, timing, and frequency of calibration to improve measurement reliability, reproducibility, and comparability.

It is worth noting that different reference body sites and reference tools themselves have varying levels of accuracy and reliability. In this systematic review, the results show that the tympanic, axillary, skin, and oral temperatures are the most commonly used reference body temperatures. Similar to non-contact measurement devices, the accuracy and reliability of reference tools are often limited by the manufacturer’s design and manufacturing process. A difference between the measured temperature of the reference body site and the true core body temperature of humans is likely to exist. Temperatures measured in the rectum and gastrointestinal tract are less susceptible

to external environmental influences than those measured at more peripheral body sites and therefore more closely reflect the body’s internal thermal state. Indeed, while rectal, gastrointestinal, and esophageal temperatures can be more invasive and challenging to measure, they are often considered the gold standard for body temperature measurement in clinical settings [114, 115]. In determining a reference site for non-contact body temperature measurement, the balance between the convenience of selecting a certain location and the measurement accuracy is important.

The results reveal that non-contact body temperature measurements can be conducted across various body positions, but the accuracy of these measurements can vary widely depending on the reference positions used. Among the included studies, the data were too heterogeneous to allow for a consolidated summary or the establishment of a general rule of thumb regarding the relationship between the body position measured by a non-contact thermometer and its accuracy. This highlights a significant gap in the understanding of how measurement and reference locations interact to influence accuracy. Further research is essential to explore these relationships more thoroughly, as a more precise understanding could help improve standardized guidelines or best practices for non-contact temperature measurement. Furthermore, the angle and distance of the measurement device in relation to the measurement location can further influence the outcome.

The inner canthus of the eye has been proposed as a potentially suitable facial measurement site for infrared thermography because it is relatively well perfused and generally less exposed to environmental influences than more peripheral facial regions. ISO/TR 13154:2017 recommends measuring the skin region medially adjacent to the inner canthus for thermographic fever screening [65]. Therefore, the inner canthus may be a useful shell-temperature site for static fever screening, but its suitability for continuously tracking CBT remains limited. In the included study by Fernandes et al., inner-canthus temperature showed substantial differences from gastrointestinal temperature, particularly during and after exercise [95]. Similar inconsistencies have been reported when inner-canthus temperature was compared with oesophageal temperature under rest, exercise, recovery, and passive-heating conditions [116].

Further research is therefore required before inner-canthus thermography can be considered a reliable method for remote CBT monitoring under dynamic conditions.

The results show that most of the current comparative experiments on NCITs or IRT focus on indoor tests or tests in a controlled environment, with the temperature ranging from 20 to 38 °C, and there are few tests in low-temperature environments, despite the fact that the average temperature in many countries is lower than 20 °C, such as the United Kingdom [117]. Matthew J. Maley conducted a study in a climate-controlled chamber and showed that infrared cameras over-estimate skin temperature during re-warming after cold exposure [104]. This suggests that the accuracy of non-contact temperature measurement devices may be affected by temperature fluctuations and other environmental factors. This highlights that further research is needed to assess the applicability of non-contact body temperature measurement devices in lower-temperature environments, as low ambient temperature can significantly impact the physiological characteristics of participants such as vasoconstriction [43], as well as the performance of non-contact thermometers. Meanwhile, we should also pay attention to the influence of conditions such as humidity and clothing on non-contact body temperature measurement. Most of the included studies did not report the environmental humidity or the clothing of the participants. It is suggested that environmental variables should be strictly controlled and reported in future studies.

Additionally, by comparing the MD and SD of different experimental setups at various mean reference body temperatures, the results showed that the MD and SD tended to be somewhat larger at higher body temperatures. Higher body temperature is often associated with increased metabolic activity. Variations in physiological activity accompanying higher body temperatures may induce a greater variability in the measured parameters, resulting in larger MD and SD values [46, 118].

Furthermore, the results suggested a potential relation between the MD(SD) and the average age of subjects, especially for NCITs. In this comparison, for simplification, for each age category, the reported MDs and SDs were averaged directly across the relevant experimental settings without pooling or weighting by sample size. Despite variations in settings and devices, both younger groups (average age <

20 years) (averaged MD=-0.044°C, averaged SD=0.448°C) and elder groups (average age > 50 years)(averaged MD=-0.160°C, averaged SD=0.280°C) showed better accuracy levels than groups with an average age ranging from 20 to 50 years(averaged MD=-0.282°C, averaged SD=0.685°C). However, An V. Nguyen demonstrated that the differences between readings of NCITs and the oral reference temperature for older people were larger than those obtained for younger people [119]. This divergence underscores the importance of continued research within this area. Further investigations should be conducted to elucidate the underlying factors responsible for these variations and to optimize the accuracy of non-contact thermometers across all age groups. This would not only aid in a clearer understanding of the age-related nuances of temperature readings, but also guide clinicians in making more informed decisions based on NCIT measurements.

Though most of the included studies reported the gender ratio, none of them reported the difference of the accuracy between genders. Women typically have greater surface mass and higher subcutaneous fat content, leading to different thermal responses from men [120]. The study conducted by Eduardo Borba Neves showed that women have a lower skin surface temperature in trunk, upper and lower limbs than men [121]. To improve the accuracy of non-contact body temperature measurements, further studies are needed to explore in more detail the accuracy difference between genders. In addition, none of the studies included in the review specifically investigated the impact of operator experience and training on the accuracy and reliability of non-contact temperature measurement devices. This highlights a potential area for future research in understanding the importance of operator proficiency in utilizing these devices. Moreover, among the included studies, none assessed the long-term stability and consistency of non-contact temperature measurement devices. However, it is crucial to note that, without understanding their long-term performance, the reliability of these devices for continuous monitoring or repeated measurements over extended periods remains uncertain. There might be potential drifts in accuracy due to incorrect device calibration or device aging. Future studies should be aimed at evaluating the long-term consistency of these devices, especially if they are to be used for continuous or prolonged monitoring [67].

The findings from the quality assessment for this systematic review have implications for the overall conclusions that are drawn. Studies with a high risk of bias can introduce potential biases that affect the accuracy and reliability of the results. Notably, the lack of random patient selection in several studies is a potential concern. Random patient selection is crucial for rigorous research design, as it ensures the generalizability and representativeness of the findings to the broader population. The absence of random selection can introduce performance and detection bias, leading to either the over-estimation or under-estimation of the accuracy of non-contact temperature measurement devices. For example, compared to young people, the average body temperature of the elderly is relatively lower due to reduced physical activity and inefficient thermoregulatory mechanisms, so the accuracy of temperature measurement may not be the same as that for young people [122]. For large-population monitoring, more work needs to be performed to include disabled volunteers in the sample population, as it is known that, for certain disabilities, chilled extremities or lower cardiac output could influence skin temperature readings [123].

Considering these limitations, it is important to interpret the conclusions drawn from this review with caution. While NCITs and IRT offer non-invasive and efficient methods for body temperature monitoring, their performance can be influenced by various factors. To mitigate these challenges, existing protocols should be further refined and standardized to more comprehensively account for environmental conditions, individual characteristics, and potential device-related biases in non-contact temperature measurement [65, 124]. Selection of contact sensors might be preferred if clinical accuracy is required.

Regarding fever screening, the accuracy of both NCITs and IRT is greatly influenced by the chosen cutoff temperature. A study by Daniel K. Ng investigated the effect of gradually increasing the fever screening cutoff temperature of NCITs from 34.2 °C to 36.0 °C [81]. The study found that, as the cutoff temperature increased, the sensitivity of the screening test gradually decreased from 97.56% to 50.41%, while the specificity increased from 18.81% to 93.96%. A similar trend was observed in Eddie Y.-K. Ng’s study on IRT [125]. As the cutoff temperature increased from 30.8 °C to 37.0 °C, the sensitivity of the screening test gradually

decreased from 100.0% to 30.2%, while the specificity increased from 0.0% to 98.7%. These findings suggest that the choice of cutoff temperature is a critical factor in the fever screening accuracy and reliability of non-contact temperature measurement devices. In the context of the COVID-19 pandemic, many public places, such as airports and hospitals, are using non-contact temperature measurement devices for fever screening. However, the choice of the appropriate cutoff temperature is still a matter of debate, and it is important to carefully consider the trade-offs between sensitivity and specificity when choosing a cutoff temperature for fever screening. When screening for fever, high sensitivity is often required to ensure that individuals with the disease are not missed. However, high sensitivity may result in lower specificity, potentially causing false positives, and leading to unnecessary further testing or treatment. Additionally, it is worth noting that the optimal threshold for fever screening may vary due to clothing, the surrounding environment, make-up and individual differences. Furthermore, it is also important to note that fever is not always present in individuals with COVID-19. A study by Pana, Bogdan C. tested the performance of NCITs in COVID-19 screening [126]. When the cutoff temperature was set to 37.3 °C, the sensitivity of PCR-positive patients of COVID-19 was only 9.43%. Apart from body temperature, other symptoms or factors, such as exposure history, are also necessary to be considered in screening protocols.

Accurate and timely measurement of body temperature is crucial in identifying and preventing heat-related illnesses. Existing contact temperature detection devices (oral, axillary, ear, temporal, and forehead) measure temperatures that differ significantly from rectal temperatures and are inadequate for assessing hyperthermia in individuals exercising outdoors in the heat [36]. NCITs and IRT have the potential to offer non-invasive and efficient methods for monitoring body temperature in high-temperature environments, where physical contact with individuals may be impractical or uncomfortable.

However, it is important to consider the limitations and challenges associated with non-contact temperature measurement devices in, e.g., heat illness prevention. Environmental factors, such as high ambient temperature, humidity, and exposure

to direct sunlight, can influence the accuracy and reliability of temperature measurements. The performance of these devices may be affected by factors such as sweat, evaporation, and heat radiation from the body's surface. Additionally, variations in skin characteristics, such as skin color and moisture levels, may introduce further complexities in temperature measurement accuracy.

2.5 Conclusion

This systematic review synthesizes 72 different settings from 32 research articles and presents a comprehensive evaluation of non-contact infrared body temperature measurement methods compared to contact body temperature measurement tools. The findings suggest that non-contact infrared temperature measurement devices, specifically NCITs and IRT, can offer accurate and reliable temperature readings in specific settings and conditions. A total of 29 NCIT settings and 4 IRT settings have the absolute value of a mean difference < 0.3 °C, which is equal to the laboratory accuracy requirements of the ISO 80601-2-56:2017 standard. However, the inconsistency in the level of accuracy underscores the need for more research in this area.

It is also found that the selection of the reference temperature tool and body site, as well as the calibration of non-contact devices, greatly impacted their performance. Therefore, future studies are recommended to focus on improving standardized guidelines for non-contact temperature measurement, including the selection of appropriate reference standards and proper device calibration.

Notably, the quality assessment of the included studies revealed a high or unclear risk of bias in many studies for the participant selection domain, raising concerns about the overall quality of existing evidence. Consequently, future research should aim to adhere to a rigorous methodology to minimize bias and maximize the applicability of the results.

In summary, while non-contact infrared methods show promise for static screening in controlled environments, their sensitivity to external factors and inconsistent accuracy in physical activities limit their utility for continuous heat strain monitoring. Moreover, the achievable agreement of non-contact measurements is inherently constrained by the variability and context-dependence of the contact-based reference

methods against which they are evaluated. These limitations indicate that non-contact sensing alone is insufficient for the objectives of this thesis. Consequently, these limitations necessitate a shift away from direct measurement towards indirect estimation strategies based on physiological signals, which are explored in the subsequent chapters.

3

Data-Driven Estimation of Core Body Temperature during Physical Activity under Heat Stress: A Systematic Review and Standardized Evaluation

Chapter 2 demonstrated that non-contact temperature measurement methods are not suitable for continuous core body temperature (CBT) monitoring under dynamic conditions, and that contact-based or indirect approaches remain necessary. Building on this finding, this chapter focuses on data-driven CBT estimation methods that rely on wearable physiological signals.

This chapter presents a systematic review of data-driven CBT estimation approaches, together with a standardized benchmark evaluation of representative models. A wide range of modeling categories are examined, including Kalman filter or extended Kalman filter (EKF/KF) models, statistical models, machine-learning models, and hybrid models. The review and benchmark evaluation reveal variability in model performance across different datasets, activity protocols, and input configurations, and show that no single CBT estimation approach consistently outperforms others across all settings.

Overall, EKF/KF methods based solely on heart rate provide a stable baseline under limited sensing conditions, whereas multimodal approaches incorporating additional physiological signals can offer conditional performance gains in more controlled scenarios. These findings clarify the strengths and limitations of existing data-driven CBT estimation methods and provide the methodological foundation for the developments introduced in the subsequent chapters.

3.1 Introduction

Elite athletes, warfighters, and outdoor laborers increasingly operate in thermally challenging environments where sustained elevations in core body temperature (CBT) pose substantial risks to physical performance, cognitive function, and survival [9–11]. Exertional heat illness (EHI), particularly exertional heat stroke (EHS), remains among the leading causes of mortality in sport and constitutes a persistent readiness concern for military and emergency-response personnel [12, 13]. EHS is the marked elevation of core body temperature, typically exceeding 40 °C, accompanied by signs of organ system dysfunction due to hyperthermia [14, 15]. The risk of morbidity and mortality increases sharply with the duration that body temperature remains above critical thresholds, particularly beyond 41 °C, but can be significantly reduced through rapid and effective cooling [16, 17]. Accordingly, contemporary sports science and occupational physiology emphasize continuous heat-strain surveillance during training camps, competitions, and simulated operational tasks to guide individualized cooling interventions, optimized work–rest cycles, and acclimatization strategies [18–20].

To support heat-strain surveillance, practitioners often begin by assessing environmental conditions. The most widely used metric is the Wet-Bulb Globe Temperature (WBGT), which integrates air temperature, humidity, solar radiation, and wind speed into a single index of external heat load [21, 22]. WBGT helps guide training modification and event scheduling by identifying conditions under which heat strain is likely to occur [23, 24]. However, WBGT reflects only environmental heat stress and cannot account for key determinants of internal heat strain, such as metabolic rate, exercise intensity, heat-acclimation status, individual tolerance, or clothing and equipment burden. It also fails to capture rapid within-session fluctuations in thermal load or indicate when an athlete is approaching a dangerous rise in core body temperature [25, 26]. As a result, heat-related illnesses may still occur even under WBGT-guided management and extensive preventive measures, as observed in large-scale sporting events such as the Tokyo 2020 Olympic and Paralympic Games [27, 28]. Thus, while WBGT is valuable for environmental risk classification, it cannot replace physiologically grounded monitoring capable of detecting changes in internal temperature in real time.

In addition to environmental indices, heat-strain management in both athletic and occupational settings commonly incorporates physiological safety principles aimed at maintaining thermal equilibrium during prolonged activity. In practice, work–rest scheduling and exposure limits are often defined to prevent excessive internal heat storage and to avoid sustained elevations in core body temperature beyond approximately 38°C over extended periods [127–129].

Unlike WBGT, which reflects only external thermal conditions, these management strategies explicitly target the control of internal thermal strain. This distinction reinforces the importance of knowing an individual’s actual core body temperature during exercise, particularly under hot and humid conditions where internal heat load can rise rapidly and unpredictably. Consequently, accurate and preferably real-time assessment of CBT remains essential for effective heat-strain mitigation and the prevention of exertional heat illness.

In practice, however, monitoring internal temperature remains challenging. Traditional thermometry measurements such as oral, axillary, and tympanic, are convenient in clinical or resting conditions but do not reliably track CBT during dynamic physical activity due to airflow, sweat evaporation, head movement, and variable peripheral blood flow [2, 31, 35]. Their agreement with true CBT deteriorates particularly during exercise-induced thermal transients, making them unsuitable for real-time heat-strain management [36–38]. For this reason, rectal and gastrointestinal temperatures are widely regarded as the laboratory gold standards for CBT assessment, and the temperature obtained from ingestible telemetry pills is commonly used in field-based studies [15, 29]. Yet all of these gold-standard approaches are invasive, burdensome, and unsuitable for many sporting environments: rectal probes are uncomfortable and impractical during movement; ingestible pills require repeated dosing, have variable transit times, and rely on dedicated receivers [31, 32]. Even promising engineering alternatives such as zero-heat-flux (ZHF) thermometry require specialized heated sensors, rigid fixation, and consistent thermal contact [39, 40]. These conditions are difficult to maintain during vigorous movement, heavy sweating, or prolonged outdoor exercise. Thus, although these methods offer high accuracy,

their reliance on specialized hardware and restrictive usage conditions limits their feasibility for routine athlete monitoring.

Beyond direct measurement approaches, a parallel line of research has explored fully mechanistic thermophysiological models that predict CBT from fundamental principles of heat production, storage, and transfer [130–132]. These models typically represent the body as a system of one or more thermal compartments and express metabolic heat generation, conductive and convective heat exchange, blood flow-mediated heat transport, and evaporative cooling through coupled differential equations. However, these models require numerous physiological parameters such as metabolic rate, perfusion rate, and workload that are difficult to measure or estimate in real time, and the high computational demands of some finite-element models make them impractical for deployment on wearable devices [133, 134]. Consequently, despite their strong physiological grounding, purely mechanistic models have seen limited application in real-world sports and occupational monitoring, where low-latency and individualized estimates of CBT are required.

Over the past decades, substantial research has sought to bridge this monitoring gap through data-driven algorithms capable of inferring CBT from wearable sensor data and environmental inputs. Early studies relied on statistical regressions incorporating multiple inputs, such as skin temperature at varied locations, heat flux, and environmental conditions [49, 135, 136]. These models provided valuable empirical insight into human thermoregulation; however, their static functional forms inherently limited their ability to capture the dynamics arising from changes in workload or environmental conditions and restricted their generalizability across individuals. Moreover, these models frequently relied on multiple peripheral inputs, which further constrained their feasibility for real-world wearable deployment. Subsequent applications of state-space modeling and Kalman filtering enabled the fusion of sequential heart-rate histories, skin temperatures, and individualized priors into adaptive estimators [137–139]. Seminal work by Buller and colleagues demonstrated that heart-rate driven models could maintain tracking accuracy within approximately 0.2–0.3 °C root mean square error (RMSE) during real-world walking, agricultural

work, and military ruck marches, inspiring interest in embedding such algorithms into wearable devices and real-time command platforms [46, 140, 141].

The methodological landscape has since diversified substantially. Existing non-invasive CBT estimation approaches broadly span statistical regression models, Kalman filter or extended Kalman filter models, machine learning methods, and hybrid formulations that integrate biophysical structure with data-driven components [142–145]. Reported performance below 0.2 °C RMSE is increasingly common. However, these results largely stem from carefully controlled laboratory or heat chamber experiments, where ground-truth CBT is reliably captured using ingestible pills or rectal probes and environmental conditions are precisely controlled. Despite promising progress, critical challenges remain; algorithm transparency varies widely, independent validation cohorts are scarce, comparisons across model classes are hindered by inconsistent reporting of input types, sampling rates and sensor placements. Without standardized evaluation frameworks, practitioners cannot fairly benchmark models, replicate findings, or integrate algorithms into athlete monitoring systems and wearable platforms.

Two prior reviews have attempted to summarize this evolving field. Falcone et al. focused on occupational CBT prediction and highlighted the absence of uniform, model-to-model comparisons, particularly under protective clothing or heat-stress conditions [146]. Dolson et al. surveyed wearable technologies for CBT estimation in athletic and clinical applications, underscoring the promise of data-driven monitoring but noting substantial methodological heterogeneity and limited external validation [147]. Since their publication, further work—including refinements of regression-based and Kalman filter approaches as well as emerging machine-learning and deep-learning models—has broadened the available evidence [148, 149]. These developments underscore the need for an updated synthesis focused on sports, exercise, and high-performance heat-stress applications.

To address these gaps, the present work focuses specifically on data-driven estimation of core body temperature during physical activity under heat stress and combines a comprehensive systematic review with a standardized evaluation framework. The systematic review synthesizes how wearable-derived physiological

signals and environmental measurements have been exploited for CBT estimation, and systematically summarizes the reported algorithmic methods, input signal configurations, and model performance across sports, occupational, and military applications. The standardized evaluation benchmarks eligible data-driven algorithms using two previously published datasets involving structured treadmill walking and staged cycling protocols, with rectal temperature as the reference CBT and heart rate and multi-site skin temperatures as the non-invasive inputs [150–152]. Together, these components provide the first integrated synthesis and quantitative comparison of data-driven CBT estimation methods, offering a unified reference for future algorithm development, validation, and deployment in real-world heat-strain monitoring.

3.2 Methods

3.2.1 Systematic Review

Review Framework

This systematic review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [153]. The review aimed to identify real-time data-driven methods for estimating core body temperature (CBT) during exercise, occupational heat stress, military tasks, or other dynamic activities using physiological signals obtainable from wearable or non-invasive sensors.

The objectives were to (1) characterize the modeling paradigms used in CBT estimation, (2) summarize the physiological and environmental inputs across studies, as well as the ground-truth CBT measurement methods, (3) synthesize and compare the validation metrics reported across studies, and (4) identify methods suitable for reproducible implementation in the standardized benchmarking analysis.

Search Strategy

A structured literature search was conducted in *Web of Science*, *PubMed*, and *IEEE Xplore* from database inception to June 2025. The following Boolean keyword query was applied systematically across all databases, with database-specific syntax adapted as needed:

*Human AND (“core body temperature” OR “deep body temperature”
OR “core temperature” OR “rectal temperature”) AND (estimation OR
prediction OR model)*

The keywords were selected to cover the study population, the target temperature variable, and the modelling task. Broad terms and truncation were used to maximize the retrieval of potentially relevant studies, with final relevance determined during study screening.

All retrieved records were imported into EndNote for reference management and deduplication. In addition, the reference lists of included articles and relevant review papers were manually screened to identify further eligible studies.

Eligibility Criteria

Eligibility criteria were defined to identify studies relevant to real-time, data-driven estimation of CBT during physical activity. Studies were eligible if they:

1. developed or evaluated a method for predicting or estimating human CBT with sufficient methodological detail for interpretation;
2. employed a data-driven modeling approach (e.g., statistical models, Kalman or extended Kalman filtering, machine learning, deep learning, or hybrid methods);
3. used physiological or environmental inputs obtainable in real time from wearable or otherwise non-invasive sensors (e.g., heart rate, skin temperature, heat flux);
4. collected data during dynamic physical activity (e.g., exercise, occupational work, military tasks);
5. used a valid reference CBT measurement (rectal, esophageal, or gastrointestinal telemetry pill);
6. were published as peer-reviewed journal articles or full conference papers in English.

Study Screening

Screening followed the PRISMA workflow. Titles and abstracts were independently screened by Zhao and Li to remove duplicates, irrelevant records, and non-data-driven studies. Full text articles were then assessed in detail against all eligibility criteria. Any disagreements were resolved through discussion, and when necessary, adjudicated by Bergmann. The PRISMA flow diagram summarizing identification, screening, exclusion, and inclusion is shown in Fig. 3.1.

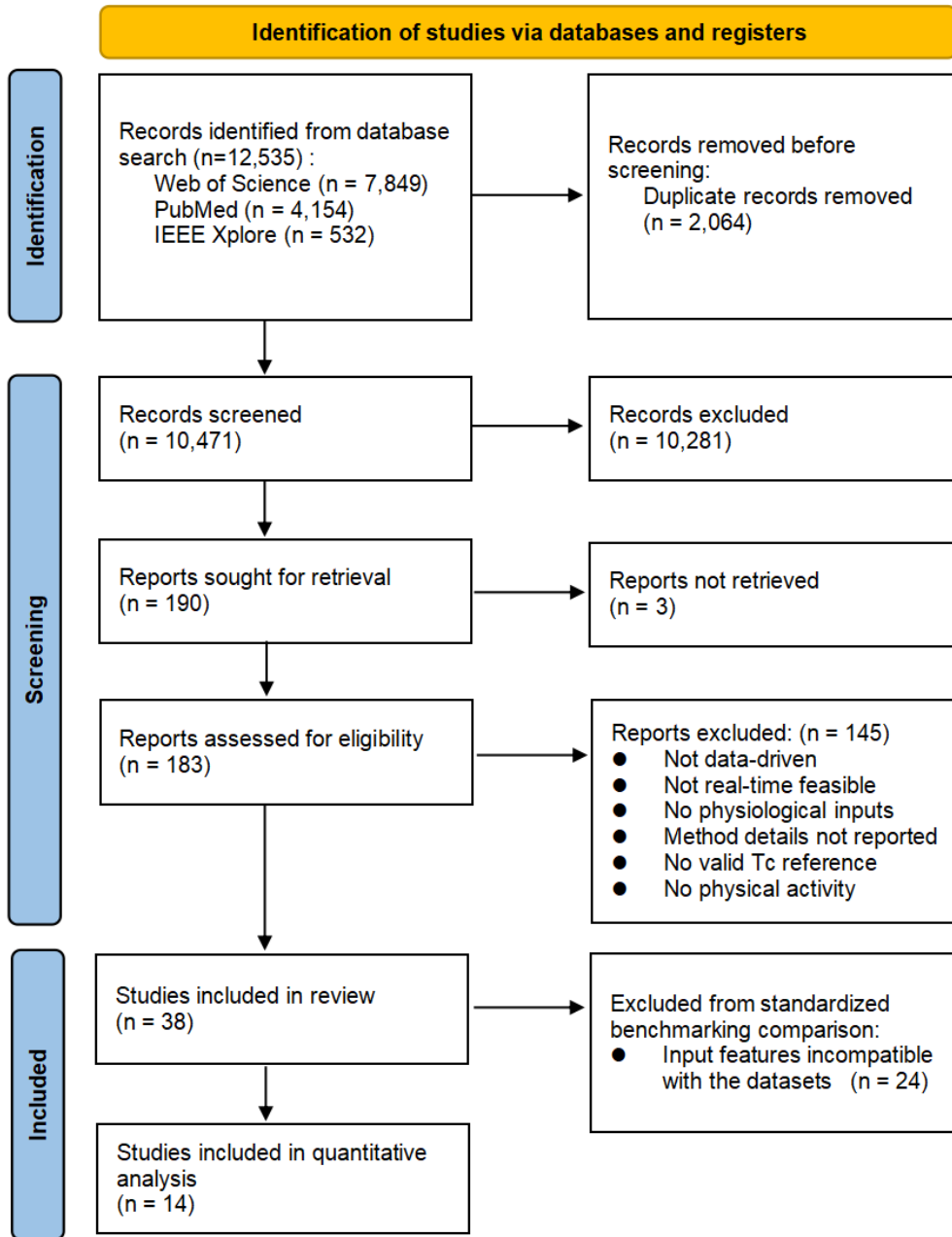


Figure 3.1: PRISMA flow diagram illustrating the identification, screening, eligibility assessment, and inclusion of studies in the systematic review.

Data Extraction

A structured extraction form was used to capture key methodological and experimental information from each included study. Extracted data included participant and experimental characteristics such as participant type (e.g., athletes, military personnel, occupational workers), sample size, sex distribution, age, body mass, and height. Environmental conditions (ambient temperature and relative humidity) and clothing or protective equipment were recorded when available. Exercise protocols were documented in detail, including the type of activity (e.g., walking, cycling), prescribed intensity (e.g., speed, workload, target heart rate), duration, and whether efforts were continuous or intermittent with rest or recovery periods. The duration of continuous monitoring was also extracted.

Information on measurement devices included the specific wearable or non-invasive sensors used (e.g., heart rate, skin temperature, heat flux), the reference CBT method (rectal, esophageal, or gastrointestinal telemetry pill), and the sampling rate of physiological measurements. Model-related information comprised the modeling approach (e.g., statistical models, Kalman or extended Kalman filtering, machine learning, deep learning, or hybrid methods), input variables, and implementation details when provided.

Extracted performance metrics included the root mean square error (RMSE), mean absolute error (MAE), Bland–Altman mean difference (MD) and standard deviation (SD), and the coefficient of determination (R^2). Many included studies evaluated multiple modeling approaches or reported several variants of the same model (e.g., identical algorithmic architecture evaluated with different input configurations). To ensure consistent comparison across publications and to avoid overweighting studies that tested a large number of model variants, a predefined extraction rule was applied: (1) For each modeling category within a study (e.g., EKF/KF, statistical, machine learning, hybrid), only the primary or best-performing model variant was extracted for inclusion in the summary tables and comparative analyses. (2) When a study reported multiple input configurations for the same primary model (e.g., heart rate only; heart rate and three skin-temperature sites; heart rate and full multi-site skin temperature), all reported configurations and their associated performance

were recorded. These details were retained because input composition substantially influences data requirements and practical deployability. This extraction strategy ensures fairness across studies, avoids inflating the influence of publications that evaluated numerous model permutations, and maintains transparency regarding the relationship between model performance and input availability.

A full expanded table containing all extracted study characteristics (including participant profiles, experimental protocols, sensor configurations, input variables, and reported performance metrics) is provided in the supplementary materials.

3.2.2 Standardized Benchmarking

To enable fair, head-to-head comparison of data-driven core body temperature (CBT) estimation models identified in the systematic review, a standardized benchmarking framework was implemented using two controlled experimental datasets.

Datasets

Dataset 1 comprises treadmill-based heat-exposure trials conducted under six controlled environmental settings [150, 151]. The settings varied in ambient temperature (25–40°C), relative humidity (20–50%), solar radiation (0 or 530 W·m⁻²), and clothing type (permeable or impermeable). In addition to heart rate and rectal temperature, the dataset includes a broad range of physiological measurements. These include multi-site skin temperatures recorded at eleven anatomical locations (abdomen, calf, chest, foot, hand, head, lower arm, lower back, thigh, upper arm, and upper back). Several experiment-specific variables such as insulated skin temperatures, insulated climatic temperatures, environmental conditions and breathing rate were also collected in the original study but were not used in the benchmarking analysis. Each trial consisted of alternating rest and treadmill walking phases, with cooling periods introduced until rectal temperature decreased by 0.4°C before subsequent exercise bouts.

Trials with missing data in any variable required for the benchmarking analysis were excluded from the dataset. After applying these completeness criteria, the six environmental settings included 3, 4, 9, 5, 10, and 4 complete trials, respectively,

and a total of 16 participants (8 males and 8 females) contributed at least one complete trial (Table 3.1).

Dataset 2 includes 13 male participants aged 18 to 45 years [152]. Each participant completed two heat-stress stages designed to impose different thermal load conditions. In each stage, participants first rested outdoors for 15 minutes, then indoors for another 15 minutes, followed by 20–60 minutes of indoor cycling performed under controlled heat stress. Exercise was terminated upon meeting one of the following criteria: a rectal temperature increase of at least 1.5°C above baseline, rectal temperature exceeding 38.5°C , voluntary exhaustion, or completion of 75 minutes of cycling. A final 30-minute outdoor rest period followed.

In Stage 1, participants wore light athletic clothing and cycled at 75% of maximal heart rate, whereas in Stage 2 they wore firefighter protective gear and cycled at 50% of HR_{max} . A comprehensive set of physiological and environmental variables was collected throughout the protocol. These included heart rate, rectal temperature, and multi-site skin temperatures recorded at several anatomical locations. The original study additionally measured insulated skin temperatures and local skin heat-flux using specialized thermistors and heat-flux sensors; however, these experiment-specific measurements were excluded from the present benchmarking analysis to maintain consistency with Dataset 1 and to ensure reliance on universally obtainable, non-insulated skin-temperature signals. Participant characteristics and environmental conditions are summarized in Table 3.2.

In both datasets, heart-rate values were provided directly as outputs from chest electrocardiogram (ECG) measurement systems. HR was not re-derived from the raw ECG recordings in this thesis, and no additional denoising or smoothing procedures were applied. The device-reported dynamic HR series were used at the original temporal resolution of each dataset.

A schematic illustration of the measurement sites used for benchmarking in Dataset 1 and Dataset 2 is provided in Fig. 3.2.

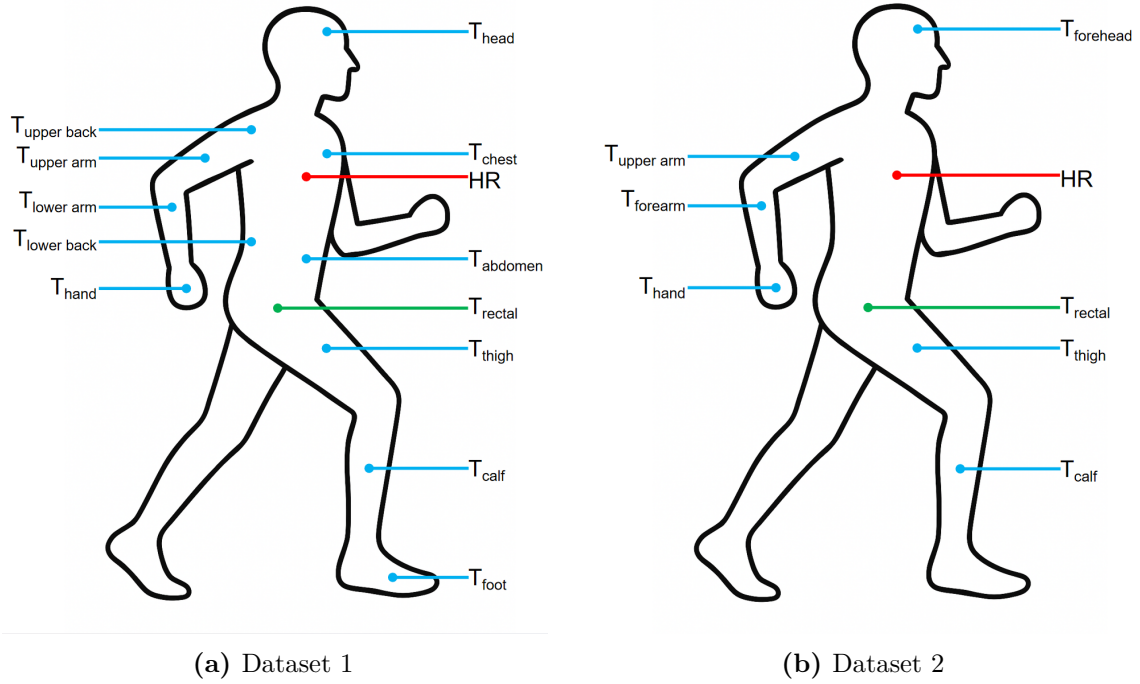


Figure 3.2: Schematic illustration of the anatomical measurement sites used for benchmarking in Dataset 1 and Dataset 2. Only the signals included in the benchmarking analysis are shown.

Table 3.1: Summary of participant demographics and experimental settings of Dataset 1.

Attribute	Value
Number of participants	16 (8 males / 8 females)
Mean age ($\pm$ SD)	25.1 $\pm$ 5.6 years
Mean body mass ($\pm$ SD)	69.3 $\pm$ 10.8 kg
Mean height ($\pm$ SD)	174.0 $\pm$ 7.3 cm
Mean body fat ($\pm$ SD)	17.1 $\pm$ 6.1%
Clothing conditions	Permeable or impermeable coveralls
Environmental settings	Setting 1: 25°C, 50% RH, impermeable, no solar Setting 2: 40°C, 20% RH, permeable, no solar Setting 3: 40°C, 20% RH, impermeable, no solar Setting 4: 30°C, 35% RH, permeable, solar (530 W·m ⁻² on back) Setting 5: 40°C, 20% RH, permeable, solar (530 W·m ⁻² on back) Setting 6: 40°C, 35% RH, permeable, no solar
Activity protocol	Setting 1: 10 min rest → 40 min treadmill walking → rest (22°C, 50% RH) until rectal temperature drops by 0.4°C → 40 min treadmill walking. Setting 6: 10 min rest in climatic chamber → 60 min treadmill walking → rest (22°C, 50% RH) until 0.4°C drop in rectal temperature → 40 min treadmill walking. Setting 2–5: Same protocol as Setting 1.
Signals used for benchmarking	Heart rate (ECG chest based Equivital EQ-02+ LifeMonitor, Hidalgo, UK), rectal temperature(rectal probe, Edale Instruments, UK), multi-site skin temperatures measured at the abdomen, calf, chest, foot, hand, head, lower arm, lower back, thigh, upper arm, and upper back (iButtons DS1922L, Maxim Integrated, US).
Sampling interval	1 minute

Table 3.2: Summary of participant demographics and experimental settings of Dataset 2.

Attribute	Value
Number of participants	13 (13 males)
Mean age ($\pm$ SD)	30.9 ± 5.4 years
Mean body mass ($\pm$ SD)	77.5 ± 6.1 kg
Mean height ($\pm$ SD)	179.2 ± 6.4 cm
Mean body fat ($\pm$ SD)	$13.1 \pm 4.3\%$
Clothing conditions	T-shirts and shorts; firefighter protective gear
Environmental settings	35°C, 57% RH, no solar
Activity protocol	15 min rest (23°C, 25% RH) $\rightarrow$ 15 min rest (35°C, 57% RH) $\rightarrow$ 20–60 min cycling (35°C, 57% RH) $\rightarrow$ 30 min rest (23°C, 25% RH)
Signals used for benchmarking	Heart rate(ECG chest belt, Unico Swiss Tex, Switzerland), rectal temperature(rectal thermometer, MSR Electronics, Switzerland), multi-site skin temperatures measured at the head, upper arm, forearm, hand, thigh, and calf (iButtons DS1922L, Maxim Integrated, US).
Sampling interval	10 seconds

Benchmark Input Signals

To ensure consistent and interpretable model comparison, all CBT-estimation algorithms were evaluated using a standardized set of input signals selected from each dataset. Only variables that were consistently available, non-invasive, and commonly obtainable in exercise or field monitoring contexts were included.

For Dataset 1, the benchmarking inputs consisted of heart rate and multi-site skin temperatures recorded at eleven anatomical locations (abdomen, calf, chest, foot, hand, head, lower arm, lower back, thigh, upper arm, and upper back). For Dataset 2, the benchmarking inputs included heart rate and multi-site skin temperatures measured at six anatomical locations (head, upper arm, forearm, hand, thigh, calf).

Rectal temperature served exclusively as the ground-truth reference in both datasets and was not used as a model input. Experiment-specific measurements—such as insulated skin temperatures and insulated climatic temperatures in Dataset 1, and heat-flux signals in Dataset 2—were excluded to maintain comparability across models and ensure practical relevance for real-world deployment.

Tables 3.1 and 3.2 list the exact benchmarking signals used for each dataset.

Evaluation Framework

A standardized evaluation framework was implemented to ensure fair comparison across all CBT estimation models considered in the benchmarking analysis. All signals

used for benchmarking were used in their recorded form without additional smoothing, filtering, or interpolation, apart from the intrinsic synchronization performed during data acquisition. This ensured that no dataset-specific preprocessing biased the comparison between methods.

To provide a comprehensive assessment of model performance across representative heat-stress scenarios, all CBT-estimation methods were evaluated independently on Dataset 1 and Dataset 2. Each dataset constitutes a separate benchmarking environment, and all comparisons were performed strictly within-dataset. This design yields two parallel evaluations—one for each dataset—ensuring that every model is assessed under identical conditions for a given dataset and allowing consistent, head-to-head comparison across modeling approaches.

For all trainable models—including regression-based approaches, Kalman-filter variants, and machine-learning architectures—all parameters were re-estimated exclusively using the training folds of each dataset. Published coefficients from prior studies were not used, as those parameters were obtained under different populations, sensor configurations, and environmental conditions, and would therefore introduce dataset-specific bias and compromise the fairness of the benchmark.

A three-fold, trial-based cross-validation scheme was adopted. All complete trials from Dataset 1 and Dataset 2 were treated as independent units and partitioned into three mutually exclusive folds of approximately equal size. Trials were partitioned using a purely trial-based three-fold cross-validation scheme. Due to the limited number of complete trials per participant, it was not feasible to assign all trials from the same participant to a single fold. Consequently, trials from the same participant could appear in different folds. This design was necessary to ensure that each fold contained a sufficient number and diversity of trials for both training and testing. In each cross-validation iteration, two folds were used for model training (and parameter estimation when required), while the remaining fold served as the held-out test set. This procedure was repeated three times such that each fold acted once as the test set.

For models that required a separate validation set (for example, to tune hyperparameters or implement early stopping), an internal validation subset was defined within the training data only. Specifically, a small proportion of the training trials

(approximately 25%) was reserved as a validation set, while the remaining training trials were used for parameter fitting. The outer three-fold cross-validation structure and the test folds remained unchanged, ensuring that test data were never used for model selection or tuning.

All models were executed in a causal, forward-in-time manner to preserve real-time feasibility. Predictions were generated sequentially without access to future samples, and no retrospective smoothing or bidirectional filtering was applied. Because different models require different warm-up periods (e.g., minimum sequence length for recurrent architectures), performance metrics were computed only over the temporal intervals in which all models produced valid outputs. For each trial, the evaluation window was aligned by discarding the initial samples up to the maximum warm-up length across all models, ensuring that every model was evaluated over an identical, causally valid time segment. Ground-truth CBT and model-predicted CBT were then aligned sample-by-sample within this common evaluation window.

All benchmarking models were implemented and evaluated using Python 3.10, NumPy 2.0.2, SciPy 1.13.1, scikit-learn 1.5.2, and TensorFlow 2.20.0. The analyses were conducted on a consumer laptop equipped with an AMD Ryzen 7 processor without GPU acceleration.

Model Inclusion Criteria for Benchmarking

While the systematic review identified a wide range of core body temperature (CBT) estimation approaches, only a subset was eligible for inclusion in the standardized benchmarking analysis. This distinction between reviewed models and benchmarked models was necessary to ensure methodological consistency, fair comparison, and practical relevance for real-world deployment.

Models were included in the benchmarking comparison if they satisfied two feasibility criteria:

First, the model must rely exclusively on signals available in the standardized datasets used for benchmarking, namely heart rate, non-invasive multi-site skin temperatures, and standard environmental parameters such as ambient temperature and relative humidity. Algorithms requiring experiment-specific or non-field-deployable measurements—for example heat flux, insulated skin temperature, or metabolic

rate—were excluded to maintain a like-for-like comparison across methods and to reflect the constraints of wearable monitoring in sports and occupational settings.

Second, the original publication must provide a sufficiently detailed mathematical formulation or algorithmic description to permit re-implementation using the available inputs.

As a result, several studies identified in the systematic review—particularly those relying on heat flux or insulated skin temperature—were not included in the quantitative benchmarking analysis due to input incompatibility. The benchmarking comparison therefore focuses on algorithms that can be implemented using cardiovascular and thermometric signals commonly obtainable in exercise and heat-stress monitoring scenarios.

Performance Metrics

Model performance was evaluated using multiple complementary metrics that capture pointwise accuracy, agreement with the reference measurement, temporal tracking fidelity, and heat-strain detection capability. Pointwise accuracy was quantified using the root mean square error (RMSE) and mean absolute error (MAE). Agreement between predicted and reference CBT was assessed using Bland–Altman statistics, including the mean difference (MD) and standard deviation of differences (SD). The coefficient of determination (R^2) was used to quantify the proportion of variance explained by each model.

To assess the ability to identify physiologically meaningful elevations in CBT, a threshold of 38.0°C was applied in accordance with occupational heat-stress guidelines [9]. Binary classification performance was then assessed using the area under the ROC curve (AUC), accuracy, sensitivity, and specificity. All performance metrics were computed strictly within the aligned evaluation window described above.

3.3 Results

The database search identified a total of 12,535 records. After removal of duplicates, title and abstract screening, 183 full-text articles were assessed for eligibility. Of these, 38 studies met all inclusion criteria for the systematic review.

The PRISMA flow diagram in Fig. 3.1 summarizes the identification, screening, exclusion, and final inclusion process.

Among the 38 included studies, 24 were excluded from the standardized benchmark analysis because their algorithms required incompatible inputs (e.g., heat flux, insulated skin temperature, metabolic rate, or detailed clothing insulation parameters). These excluded studies are listed in Table 3.3.

The remaining 14 studies satisfied all feasibility requirements for benchmarking. These models relied solely on signals available in Dataset 1 and Dataset 2 (i.e., heart rate, non-insulated multi-site skin temperatures, and standard environmental parameters) and provided sufficient methodological detail to allow re-implementation within the standardized evaluation framework. The benchmark-eligible models are summarized in Table 3.4.

Complete study characteristic summaries, including the detailed quality-related descriptive information (e.g., participant profiles, experimental settings, sensor configurations), are provided in the supplementary materials.

Table 3.3: Summary of studies included in the systematic review but excluded from the standardized benchmarking analysis.

Study	Model type	Input variables	Reference method	CBT	Reported metrics
Nazarian (2021) [143]	Kalman filter	Wrist air temperature; wrist skin temperature; humidity; heart rate	Gastrointestinal temperature		RMSE: 0.13 °C; MAE: 0.10 °C; MD: 0.008 °C; SD: 0.27 °C; R^2 : 0.81
Moyen (2021) [154]	Hybrid model (machine-learning-based feature modeling + extended Kalman filter)	Heart rate; skin and ambient temperature(upper arm); skin and ambient humidity(upper arm); step rate; anthropometric data (age, height, body mass, sex)	Rectal temperature and gastrointestinal temperature		Rectal: RMSE: 0.30 °C; MAE: 0.25 °C; MD: 0.07 °C; SD: 0.32 °C.
Oleng' (2005) [155]	Hybrid model (first-principles physiology model + autoregressive (AR) error-correction)	Core temperature history; metabolic rate; environmental conditions (air temperature, humidity, wind, radiant temperature); clothing insulation and thermal properties; body geometry and physical constants.	Rectal temperature		Gastrointestinal: RMSE: 0.29 °C; MAE: 0.26 °C; MD: 0.07 °C; SD: 0.32 °C. NR
Gribok (2008) [156]	Hybrid model (first-principles physiology model + autoregressive (AR) residual correction)	Core temperature history; metabolic rate; environmental conditions; clothing insulation parameters; physical constants	Rectal temperature (laboratory study); gastrointestinal temperature (field study)		20-min prediction RMSE (same-subject): 0.22 ± 0.04 °C; (cross-subject): 0.23 ± 0.01 °C
Hamatani (2015) [157]	Hybrid model (thermoregulation model (two-node) with parameter filtering)	Wrist skin temperature; heart rate; ambient temperature; humidity; solar radiation; clothing insulation; body anthropometrics	Rectal temperature		MD: -0.07 °C
Laxminarayan (2018) [145]	Hybrid model (physics-based energy-balance + adaptive Kalman filter)	Heart rate; mean skin temperature; activity; ambient temperature; relative humidity	Rectal temperature		RMSE: 0.33 °C (overall)
Tanaka (1979) [158]	Statistical model (linear regression)	Heart rate; exercise/recovery state	Rectal temperature		NR
Gibson (1981) [159]	Statistical model (linear regression with rate-of-change term)	Auditory canal temperature and its rate of change (dT_c/dt)	Oesophageal temperature		RMSE: 0.11 °C
Hansen (1996) [160]	Statistical model (multiple linear regression)	Tympanic temperature; aural canal temperature; cheek skin temperature	Rectal temperature		RMSE: 0.15 °C (warm); 0.21 °C (hot)
Gribok (2006) [161]	Statistical model (autoregressive with exogenous inputs)	Delayed core temperature; age; height; weight; body fat %; walking speed; mean radiant temperature; relative humidity	Rectal temperature		RMSE: ≈ 0.08 °C (subject-specific); ≈ 0.19 °C (cross-subject)
Gribok (2013) [162]	Statistical model (functional regression)	Heart rate; respiratory rate; skin temperature (not specified); subcutaneous glucose concentration	Gastrointestinal temperature		NR
Xu (2013) [135]	Statistical model (multiple linear regression)	Skin temperature and skin heat flux at six sites (forehead, sternum, pectoralis, rib, scapula, thigh)	Rectal temperature		R^2 : 0.40 (forehead); 0.75 (sternum); 0.69 (scapula); 0.60 (pectoralis); 0.70 (rib); 0.59 (thigh); 0.79 (sternum + scapula + rib)
Richmond (2013) [136]	Statistical model (multiple linear regression)	Insulated neck skin temperature; microclimate temperature	Rectal temperature		RMSE: 0.20 °C; R^2 : 0.85

Study	Model type	Input variables	Reference method	CBT	Reported metrics
Niedermann (2014) [49]	Statistical model (multiple linear regression + PCA feature extraction)	Skin temperatures (upper arm, lower arm, thigh); skin heat flux (chest, back); heart rate	Gastrointestinal temperature	tem-	RMSE: 0.28 °C (no radiation), 0.28 ± 0.03 °C; 0.34 °C (with radiation), 0.34 ± 0.07 °C; MD: 0.04 ± 0.17 °C (no radiation), -0.17 ± 0.14 °C (with radiation); SD: 0.17 °C (no radiation), 0.14 °C (with radiation); R^2 : 0.72 (development), 0.70–0.73 (validation)
Richmond (2015) [150]	Statistical model (multiple linear regression)	Insulated neck skin temperature; microclimate temperature; heart rate (HR); work/rest indicator	Rectal temperature		RMSE: 0.24 °C; R^2 : 0.82
Seng (2016) [163]	Statistical model (nonlinear mixed-effects model)	Chest, upper arm, thigh, and calf skin temperature; load carried (kg); work–rest schedule; optional baseline core temperature data for updating	Gastrointestinal temperature	tem-	Population model: RMSE: 0.67 °C; MD: -0.05 °C; SD: 0.34 °C. Mixed-effects Model 1 (Tc baseline only): RMSE: 0.57 °C; MD: 0.07 °C; SD: 0.29 °C. Mixed-effects Model 2 (Tc baseline + Tc end-of-work): RMSE: 0.29 °C; MD: 0.03 °C; SD: 0.15 °C R^2 : 0.625
Gibson (2017) [164]	Statistical model (multiple linear regression)	Relative power; rating of perceived exertion; other exercise prescription variables	Rectal temperature		
Nakada (2017) [165]	Statistical model (multiple linear regression)	Three auditory canal temperatures	Oesophageal temperature		RMSE: 0.113 °C; R^2 : 0.904
Eggenberger (2018) [152]	Statistical model (PCA + multiple linear regression)	Heart rate; heat flux at sternum, scapula, and rib; insulated skin temperatures at sternum, scapula, rib, and radial artery; non-insulated skin temperatures at sternum, scapula, rib, radial artery, forehead, upper arm, forearm, hand, thigh, and calf	Rectal temperature		Max-input model (18 inputs): RMSE: 0.28 °C; R^2 : 0.703. Min-input model (2 inputs): RMSE: 0.29 °C; R^2 : 0.677
Tsadok (2021) [166]	Statistical model (curve fitting / regression models)	Core dual-sensor heat-flux temperatures (forehead and wrist)	Rectal temperature		RMSE: 0.29 °C (forehead, overall), 0.23 ± 0.10 °C (forehead, by subject); 0.40 °C (wrist, overall), 0.31 ± 0.14 °C (wrist, by subject); MAE: 0.20 ± 0.16 °C (forehead), 0.27 ± 0.20 °C (wrist); MD: 0.00 ± 0.25 °C (forehead), 0.01 ± 0.32 °C (wrist); SD: 0.25 °C (forehead), 0.32 °C (wrist)
Kurosaka (2022) [142]	Statistical model (multiple linear regression)	ECG-derived Poincaré plot features from current and previous time windows; previously estimated temperature change	Rectal temperature		RMSE: 0.023 °C; SD: ± 0.13 °C; R^2 : 0.46
Li (2024) [167]	Statistical model (linear regression)	Skin temperature at multiple sites (chest, upper arm, thigh, calf); environmental parameters (air temperature, radiant temperature, humidity, air velocity)	Rectal temperature		No direct prediction accuracy metrics were reported (RMSE, MAE, MD, SD not provided).

Study	Model type	Input variables	Reference method	CBT	Reported metrics
Wang (2023) [168]	Machine learning model (random forest regression)	Sweat sodium concentration; sweat potassium concentration; heart rate; regional sweat rate	Gastrointestinal temperature	tem-	RMSE: 0.02 °C
Tan (2024) [144]	Machine learning model (random forest regression)	Aural canal temperature; external auricle temperature; heart rate; trial phase information (baseline, exercise, recovery)	Gastrointestinal temperature	tem-	RMSE: 0.27 °C; MAE: 0.20 ± 0.18 °C; MD: -0.02 ± 0.26 °C; SD: 0.26 °C

Note: NR : Not reported.

Table 3.4: Summary of studies included in the systematic review and the standardized benchmark analysis.

Study	Model type	Model details	Input variables	Reference CBT method	Reported metrics	Continuous tracking metrics (Dataset 1)	Continuous tracking metrics (Dataset 2)	High-risk temperature detection (Dataset 1)	High-risk temperature detection (Dataset 2)
Buller (2013) [46]	Extended Kalman filter	Extended Kalman filter using heart rate as a nonlinear observation of core temperature. Time update assumes a linear autoregressive process model $CT_t = CT_{t-1} + f$, $f \sim \mathcal{N}(0, \gamma^2)$. Heart rate is mapped to CT through a quadratic observation model $HR_t = b_2 CT_t^2 + b_1 CT_t + b_0 + g$, enabling EKF linearization and recursive minute-by-minute CT estimation.	Heart rate	Gastrointestinal temperature	RMSE: 0.30 ± 0.13 °C; MD: -0.03 °C; SD: 0.32 °C; R^2 : 0.71	RMSE: 0.4108 °C, MAE: 0.2943 °C, MD: -0.0520 °C, SD: 0.4075 °C, R^2 : 0.4686 .	RMSE: 0.6563 °C, MAE: 0.4746 °C, MD: 0.1388 °C, SD: 0.6415 °C, R^2 : -0.4010 .	Accuracy: 0.8186 , Sensitivity: 0.7117 , Specificity: 0.9171 , AUC: 0.9004 , RMSE: 0.5384 °C.	Accuracy: 0.7728 , Sensitivity: 0.3328 , Specificity: 0.8549 , AUC: 0.8202 , RMSE: 0.4796 °C.
Buller (2015) [169]	Extended Kalman filter	Same as Buller (2013) [46].	Heart rate	Gastrointestinal temperature	RMSE: 0.21 ± 0.11 °C; MD: 0.02 °C; SD: 0.25 °C	Same as [46]	Same as [46]	Same as [46]	Same as [46]
Hunt (2019) [170]	Extended Kalman filter	Same as Buller (2013) [46].	Heart rate	Gastrointestinal temperature	RMSE: 0.32 °C; MAE: 0.25 ± 0.20 °C; MD: 0.01 °C; SD: 0.33 °C	Same as [46]	Same as [46]	Same as [46]	Same as [46]
Buller (2020) [52]	Extended Kalman filter	Same as Buller (2013) [46].	Heart rate	Gastrointestinal temperature	RMSE: 0.37 ± 0.13 °C (hot), 0.49 ± 0.26 °C (warm); MD: -0.10 °C (hot), 0.34 °C (warm); SD: 0.37 °C (hot), 0.40 °C (warm)	Same as [46]	Same as [46]	Same as [46]	Same as [46]
Egbert (2022) [140]	Extended Kalman filter	Same as Buller (2013) [46].	Heart rate	Gastrointestinal temperature	RMSE: 0.41 °C; MAE: 0.32 °C; MD: -0.14 °C; SD: 0.39 °C	Same as [46]	Same as [46]	Same as [46]	Same as [46]
Falcone (2022) [171]	Extended Kalman filter	Same as Buller (2013) [46].	Heart rate	Gastrointestinal temperature	Not reported in the paper.	Same as [46]	Same as [46]	Same as [46]	Same as [46]
Peggen (2024) [141]	Extended Kalman filter	Same as Buller (2013) [46].	Heart rate	Gastrointestinal temperature	RMSE: 0.37 ± 0.15 °C; MD: 0.09 °C; SD: 0.22 °C	Same as [46]	Same as [46]	Same as [46]	Same as [46]

Study	Model type	Model details	Input variables	Reference CBT method	Reported metrics	Continuous tracking metrics (Dataset 1)	Continuous tracking metrics (Dataset 2)	High-risk temperature detection (Dataset 1)	High-risk temperature detection (Dataset 2)
Seng (2016) [137]	Extended Kalman filter	Nonlinear state-space EKF with heart-rate-driven dynamics and cubic skin-temperature observation. The state transition follows a nonlinear autoregressive model, $T_{c,t} = \alpha_0 + \alpha_1 T_{c,t-1} + \alpha_2 T_{c,t-1}^2 + \alpha_3 HR_{t-1} + \omega_t$, and the observation model maps chest skin temperature to core temperature via a cubic function, $T_{sk,t} = \beta_0 + \beta_1 T_{c,t} + \beta_2 T_{c,t}^2 + \beta_3 T_{c,t}^3 + \eta_t$.	Surface chest skin temperature; heart rate	Gastrointestinal temperature	RMSE: 0.29 °C; MD: 0.11 °C; SD: 0.27 °C	RMSE: 0.3269 °C; MAE: 0.2530 °C; MD: -0.0214 °C; SD: 0.3262 °C; R^2 : 0.6634	RMSE: 0.5220 °C; MAE: 0.4122 °C; MD: 0.1466 °C; SD: 0.5009 °C; R^2 : 0.1139	Accuracy: 0.8237; Sensitivity: 0.7742; Specificity: 0.8693; AUC: 0.9201; RMSE: 0.3981 °C	Accuracy: 0.7904; Sensitivity: 0.8127; Specificity: 0.7862; AUC: 0.8740; RMSE: 0.3903 °C
Rizvi (2022) [138]	Extended Kalman filter	Modified EKF with phase-dependent measurement models. Time update: $CT_t = a1 * CT_{t-1} + a0 + \varepsilon$. HR-based observation: $HR_t = b_1 CT_t^2 + b_2 CT_t + b_3 + \eta$, with separate (b_1, b_2, b_3) for exercise and recovery, selected via a 5-min moving-average HR rule.	Heart rate	Gastrointestinal temperature	RMSE: 0.16 °C; MD: -0.03 °C; SD: 0.16 °C	RMSE: 0.3952 °C; MAE: 0.3079 °C; MD: -0.0172 °C; SD: 0.3948 °C; R^2 : 0.5082	RMSE: 0.5400 °C; MAE: 0.4279 °C; MD: 0.0472 °C; SD: 0.5379 °C; R^2 : 0.0517	Accuracy: 0.8186; Sensitivity: 0.7523; Specificity: 0.8797; AUC: 0.9028; RMSE: 0.4932 °C	Accuracy: 0.8144; Sensitivity: 0.7180; Specificity: 0.8324; AUC: 0.8778; RMSE: 0.5062 °C
Falcone (2024) [148]	Extended Kalman filter	A biphasic Kalman filter (BKFB) model with separate observation functions for core-temperature increase and decrease phases. The time-update assumes $CT_t = a1 * CT_{t-1} + a0 + \omega_t$, while the measurement model maps heart rate (HR) to core temperature through phase-specific quadratic functions, $HR_t = f_{\uparrow}(CT_t) + \eta_t$ during activity and $HR_t = f_{\downarrow}(CT_t) + \eta_t$ during recovery, where $f_{\uparrow}$ and $f_{\downarrow}$ are distinct second-order polynomials. Phase switching is governed by a heart-rate threshold rule.	Heart rate	Gastrointestinal temperature	RMSE: 0.28 ± 0.12 °C	RMSE: 0.3662 °C; MAE: 0.2700 °C; MD: -0.0170 °C; SD: 0.3658 °C; R^2 : 0.5777	RMSE: 0.4921 °C; MAE: 0.3742 °C; MD: -0.0309 °C; SD: 0.4911 °C; R^2 : 0.2124	Accuracy: 0.8270; Sensitivity: 0.7373; Specificity: 0.9096; AUC: 0.9172; RMSE: 0.4808 °C	Accuracy: 0.8478; Sensitivity: 0.4567; Specificity: 0.9208; AUC: 0.8196; RMSE: 0.7072 °C
Buller (2010) [172]	Kalman filter	A linear Kalman filter was implemented with core temperature (CT) as the latent state. CT dynamics were modeled using a first-order autoregressive process: $CT_t = a1 CT_{t-1} + a0 + \omega_t$, Heart rate (HR) was incorporated through a linear observation model derived from two anchor points in the HR-CT plane: $HR_t = a CT_t + b + \nu_t$,	Heart rate	Gastrointestinal temperature	RMSE: 0.30 °C; MD: 0.02 °C; SD: 0.18 °C	RMSE: 0.4053 °C; MAE: 0.3530 °C; MD: 0.2693 °C; SD: 0.3029 °C; R^2 : 0.4827	RMSE: 0.4712 °C; MAE: 0.3914 °C; MD: 0.0039 °C; SD: 0.4712 °C; R^2 : 0.2778	Accuracy: 0.7065; Sensitivity: 0.9744; Specificity: 0.4600; AUC: 0.9215; RMSE: 0.3146 °C	Accuracy: 0.6690; Sensitivity: 0.8411; Specificity: 0.9772; AUC: 0.7486; RMSE: 0.6690 °C

Study	Model type	Model details	Input variables	Reference CBT method	Reported metrics	Continuous tracking metrics (Dataset 1)	Continuous tracking metrics (Dataset 2)	High-risk temperature detection (Dataset 1)	High-risk temperature detection (Dataset 2)
Welles (2018) [173]	Kalman filter	A linear Kalman filter was used to estimate core temperature (CT) from head skin temperature (T_{sk}) and heart rate (HR). The state-transition model followed a first-order autoregressive form: $CT_t = a CT_{t-1} + a_0 + w_t$, The observation model linearly mapped CT to the measured signals: $\mathbf{z}_t = \begin{bmatrix} T_{sk}(t) \\ HR(t) \end{bmatrix} = \mathbf{C} CT_t + \mathbf{b} + \mathbf{v}_t$,	Skin temperature (forehead, pectoralis, scapula, lower back, vastus medialis, gastrocnemius); heart rate	Gastrointestinal temperature	Thigh: RMSE(overall)=0.30°C; RMSE(by subject)=0.28±0.10°C; MD=0.04°C, SD=0.19°C. Scapula: RMSE(overall)=0.28°C; RMSE(by subject)=0.25±0.12°C; MD=-0.03°C, SD=0.15°C. Forehead: RMSE(overall)=0.30°C; RMSE(by subject)=0.28±0.12°C; MD=0.04°C, SD=0.17°C.	Thigh: RMSE: 0.3618 °C; MAE: 0.2849 °C; MD: -0.0203 °C; SD: 0.3612 °C; R^2 : 0.5877; Forehead: RMSE: 0.4489 °C; MAE: 0.3284 °C; MD: -0.0368 °C; SD: 0.4474 °C; R^2 : 0.3654	Thigh: RMSE: 0.6799 °C; MAE: 0.5440 °C; MD: 0.1557 °C; SD: 0.6618 °C; R^2 : -0.5035; Forehead: RMSE: 0.5348 °C; MAE: 0.4156 °C; MD: -0.0222 °C; SD: 0.5343 °C; R^2 : 0.0698	Thigh: Accuracy: 0.8273; Sensitivity: 0.7986; Specificity: 0.8538; AUC: 0.9023; RMSE: 0.3858 °C; Forehead: Accuracy: 0.7911; Sensitivity: 0.6792; Specificity: 0.8941; AUC: 0.8749; RMSE: 0.5864 °C	Thigh: Accuracy: 0.7506; Sensitivity: 0.7744; Specificity: 0.7461; AUC: 0.8233; RMSE: 0.5981 °C; Forehead: Accuracy: 0.7855; Sensitivity: 0.4421; Specificity: 0.8496; AUC: 0.8107; RMSE: 0.5318 °C

Study	Model type	Model details	Input variables	Reference CBT method	Reported metrics	Continuous tracking metrics (Dataset 1)	Continuous tracking metrics (Dataset 2)	High-risk temperature detection (Dataset 1)	High-risk temperature detection (Dataset 2)
Kim (2016) [174]	Statistical model (linear regression)	Rectal temperature predicted from single-site skin temperature using site-specific linear regression, $CT = aT_{\text{skin}} + b$.	Twelve skin temperatures at forehead, chest, abdomen, upper back, upper arm, forearm, hand, thigh, calf, dorsal foot, lateral foot, and toe	Rectal temperature	R^2 : 0.826 (forehead); 0.824 (chest); 0.473 (abdomen); 0.494 (back); 0.779 (upper arm); 0.753 (forearm); 0.382 (hand); 0.688 (thigh); 0.619 (calf); 0.684 (foot); 0.606 (lateral foot); 0.571 (toe)	Forehead: RMSE: 0.5701 °C; Chest: RMSE: 0.5462 °C; Abdomen: RMSE: 0.5450 °C; Upper back: RMSE: 0.5708 °C; Lower back: RMSE: 0.5638 °C; Upper arm: RMSE: 0.5068 °C; Lower arm: RMSE: 0.5264 °C; Hand: RMSE: 0.5425 °C; Thigh: RMSE: 0.4995 °C; Calf: RMSE: 0.5093 °C; Foot: RMSE: 0.4216 °C	Forehead: RMSE: 0.5107 °C; Upper arm: RMSE: 0.5444 °C; Lower arm: RMSE: 0.5521 °C; Hand: RMSE: 0.5473 °C; Thigh: RMSE: 0.5259 °C; Calf: RMSE: 0.5304 °C	Forehead: Sensitivity: 0.2989; Specificity: 0.9971; Sensitivity: 0.4953; Specificity: 0.6789; Abdomen: Sensitivity: 0.5447; Specificity: 0.6834; Upper back: Sensitivity: 0.2226; Specificity: 0.7765; Lower back: Sensitivity: 0.3271; Specificity: 0.7720; Upper arm: Sensitivity: 0.6417; Specificity: 0.6580; Lower arm: Sensitivity: 0.5610; Specificity: 0.6400; Hand: Sensitivity: 0.5622; Specificity: 0.6247; Thigh: Sensitivity: 0.6467; Specificity: 0.6620; Calf: Sensitivity: 0.6467; Specificity: 0.6710; Foot: Sensitivity: 0.8236; Specificity: 0.6394	Forehead: Sensitivity: 0.0000; Specificity: 0.9971; Upper arm: Sensitivity: 0.0000; Specificity: 1.0000; Lower arm: Sensitivity: 0.0000; Specificity: 1.0000; Hand: Sensitivity: 0.0000; Specificity: 1.0000; Calf: Sensitivity: 0.0000; Specificity: 1.0000

Study	Model type	Model details	Input variables	Reference CBT method	Reported metrics	Continuous tracking metrics (Dataset 1)	Continuous tracking metrics (Dataset 2)	High-risk temperature detection (Dataset 1)	High-risk temperature detection (Dataset 2)
Han (2025) [175]	Machine learning model (Kalman filter + deep learning, LTSF-Informer)	Hybrid framework combining an extended Kalman filter with a long-term sequence forecasting model (Informer). The EKF estimates CT from HR using $CT_t = a_1 CT_{t-1} + a_0 + f_t$ and $HR_t = b_2 CT_t^2 + b_1 CT_t + b_0 + g_t$, providing a denoised HR-derived CT feature. The Informer then predicts future CT using ProbSparse attention and encoder distillation; the combined model outputs next-step CT forecasts.	Heart rate; 7-point skin temperature (forehead, chest, forearm, hand, thigh, calf, foot); mean skin temperature	Gastrointestinal temperature	HR only: MAE: 0.17 °C; RMSE: 0.28 °C; R^2 : 0.70. HR + 3-point T_{sk} : MAE: 0.13 °C; RMSE: 0.24 °C; R^2 : 0.72. HR + 4-point T_{sk} : MAE: 0.11 °C; RMSE: 0.18 °C; R^2 : 0.83.				

Study	Model type	Model details	Input variables	Reference CBT method	Reported metrics	Continuous tracking metrics (Dataset 1)	Continuous tracking metrics (Dataset 2)	High-risk temperature detection (Dataset 1)	High-risk temperature detection (Dataset 2)
					HR + 7-point T_{sk} : MAE: 0.09 °C; RMSE: 0.12 °C; R^2 : 0.92	HR only: RMSE: 0.4107 °C; MAE: 0.3181 °C; MD: −0.0737 °C; SD: 0.4040 °C; R^2 : 0.1991; HR + 3-point Tsk: RMSE: 0.4128 °C; MAE: 0.3187 °C; MD: −0.0985 °C; SD: 0.4008 °C; R^2 : 0.1911; HR + 4-point Tsk: RMSE: 0.4110 °C; MAE: 0.3166 °C; MD: −0.0990 °C; SD: 0.3989 °C; R^2 : 0.1982; HR + 7-point Tsk: RMSE: 0.4152 °C; MAE: 0.3211 °C; MD: −0.0993 °C; SD: 0.4031 °C; R^2 : 0.1816	HR only: RMSE: 0.3354 °C; MAE: 0.2708 °C; MD: 0.0289 °C; SD: 0.3341 °C; R^2 : 0.5949	HR only: Accuracy: 0.7301; Sensitivity: 0.8080; Specificity: 0.6138; AUC: 0.7987; RMSE: 0.4483 °C; HR + 3-point Tsk: Accuracy: 0.7316; Sensitivity: 0.8856; Specificity: 0.5019; AUC: 0.7968; RMSE: 0.4652 °C; HR + 4-point Tsk: Accuracy: 0.7372; Sensitivity: 0.8881; Specificity: 0.5121; AUC: 0.8040; RMSE: 0.4627 °C; HR + 7-point Tsk: Accuracy: 0.7271; Sensitivity: 0.8856; Specificity: 0.4907; AUC: 0.7915; RMSE: 0.4672 °C	HR only: Accuracy: 0.8533; Sensitivity: 0.0000; Specificity: 1.0000; AUC: 0.8343; RMSE: 0.4898 °C

Note: Due to space constraints, for Kim (2016) [174] only RMSE, sensitivity and specificity are reported in the table. NR : Not reported; CT_t : core temperature at time t ; HR_t : heart rate at time t ; $T_{sk,t}$: skin temperature at time t ; $y(t)$: core-temperature time series used in autoregressive modeling; $\hat{C}T_{t+1}$: one-step-ahead predicted core temperature; $\tilde{C}T_t$: filtered or smoothed core-temperature signal; z_t : observation vector at time t ; f , f_t , g , g_t : random noise terms in the process or observation models; a_i , b_i , c_i : model or regression coefficients; w_t , v_t , ν_t , ε_t , η_t , ω_t : process or measurement noise terms.

3.3.1 Model categories and inputs

Across the 38 studies included in the systematic review, four modeling categories were identified: (1) Kalman filter or extended Kalman filter (KF/EKF) models, (2) statistical models, (3) pure machine-learning (ML) models, and (4) hybrid models integrating physiological or biophysical formulations with data-driven components. Statistical models were the most common (17 studies), followed by EKF/KF approaches (13 studies). Hybrid models were reported in 5 studies, and 3 studies used pure ML methods. The distribution of these four method categories is shown in Fig. 3.3(a).

Most EKF/KF-based approaches implemented the canonical heart-rate-driven extended Kalman filter in which core temperature is treated as a latent state with autoregressive time update and heart rate is linked to core temperature through a nonlinear observation function. Kalman filtering is formulated within a state-space framework consisting of a state-transition model and an observation model. A detailed description of the EKF and the state-space formulation used throughout this thesis is provided in Section 4.2.2 and Section 5.2.2. In these state-space formulations, the variable of interest (core body temperature) is represented by a hidden state that evolves over time according to a state-transition model, while sensor measurements (heart rate) are related to the hidden state through an observation model [176]. The conventional Kalman filter assumes linear state-transition and observation models, whereas the EKF extends this framework to nonlinear models by locally linearizing the nonlinear functions during the estimation process [177]. Variants of this framework included the cubic observation function in which chest skin temperature is the primary observation, phase-dependent EKFs with separate exercise and recovery observation models, and biphasic Kalman filters with distinct observation functions for heating and cooling phases. In addition to these extended Kalman filters, several studies used linear Kalman filters with either heart rate and skin temperature or core temperature history as the primary inputs.

Statistical models spanned a broad spectrum of regression and time-series formulations. Frequently used approaches included multiple linear regression linking

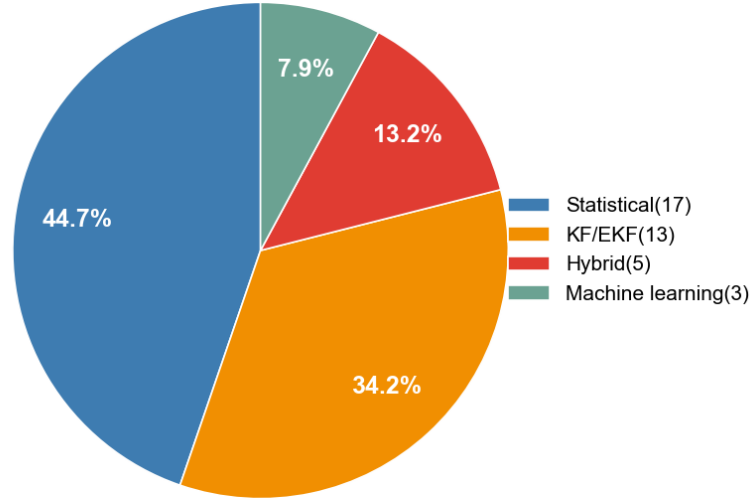
core temperature to one or more skin temperatures or tympanic/aural measurements, and regression frameworks augmented with principal component analysis. Other statistical models comprised nonlinear mixed-effects models incorporating individual baseline or end-of-work core temperature, functional regression using multiple physiological inputs, and curve-fitting models applied to dual-sensor heat flux measurements.

Pure ML models were less frequently reported and mainly consisted of ensemble and deep learning methods. Two studies employed random forest regression using combinations of cardiovascular, thermometric, and sweat composition variables, whereas one study used a long-term sequence forecasting architecture (Informer) that consumed heart rate and multi-site skin temperatures, either directly or in combination with features derived from a preceding Kalman filter. Hybrid models, in contrast, explicitly combined physics-based or thermophysiological formulations with data-driven components—for example, first-principles thermoregulation or energy balance models with autoregressive error-correction, or two-node thermoregulation models with parameter filtering and adaptive Kalman filtering.

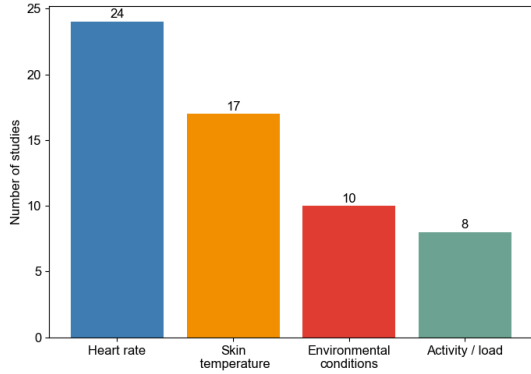
The frequency of input variables varied across studies. The five most commonly used input categories were heart rate (24 studies), skin temperature (17 studies), environmental conditions (10 studies), load- or activity-related variables (8 studies). These counts are summarized in Fig. 3.3(b).

Among studies using skin temperature, the five most frequently reported anatomical measurement sites were the chest (9 studies), thigh (8 studies), arm (7 studies), forehead (6 studies), and calf (6 studies). These distributions are illustrated in Fig. 3.3(c).

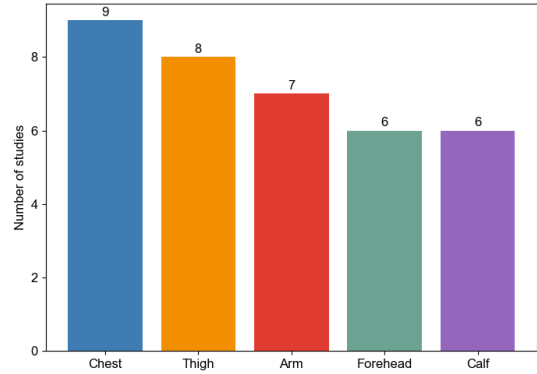
Across the included studies, validated internal core-temperature measurements were used as the reference standard. Gastrointestinal telemetry pills were the most frequently employed reference (19 studies), followed by rectal thermometry (15 studies). Two studies used oesophageal temperature, and two studies reported both rectal and gastrointestinal measurements.



(a) Model categories



(b) Input variable categories



(c) Skin-temperature measurement sites

Figure 3.3: Overview of modeling approaches and input characteristics across the 38 studies included in the systematic review. (a) Distribution of the four modeling categories: EKF/KF models, statistical models, pure ML models, and hybrid physiological-data-driven models. (b) Five most frequently used input variable categories. (c) Five most frequently reported anatomical skin-temperature sites.

3.3.2 Experiment settings

Across the 38 included studies, substantial heterogeneity was observed in participant sample size, environmental exposure, and activity protocols (Fig. 3.4).

Participant group size varied widely (Fig. 3.4(a)). A total of 13 studies enrolled ≤ 10 participants, 12 studies included 11–20 participants, 10 studies included 21–50

participants, and only 3 studies enrolled more than 50 participants.

Environmental conditions also showed wide variability (Fig. 3.4 (b)). Using three temperature bands: Cold ($<20^{\circ}\text{C}$), Moderate ($20\text{--}29.9^{\circ}\text{C}$), and Hot ($\geq 30^{\circ}\text{C}$), it was found that 6 studies included at least one cold-temperature protocol, 22 studies included moderate-temperature conditions, and 27 studies involved hot environments. A single study did not report ambient temperature and was excluded from this figure.

A similarly diverse range of activity protocols was observed (Fig. 3.4 (c)). Walking or running was the most frequently employed modality (24 studies), followed by military or load-carriage tasks (9 studies) and cycling protocols (8 studies). Because these categories are not mutually exclusive, individual studies could contribute to multiple activity types or temperature bands.

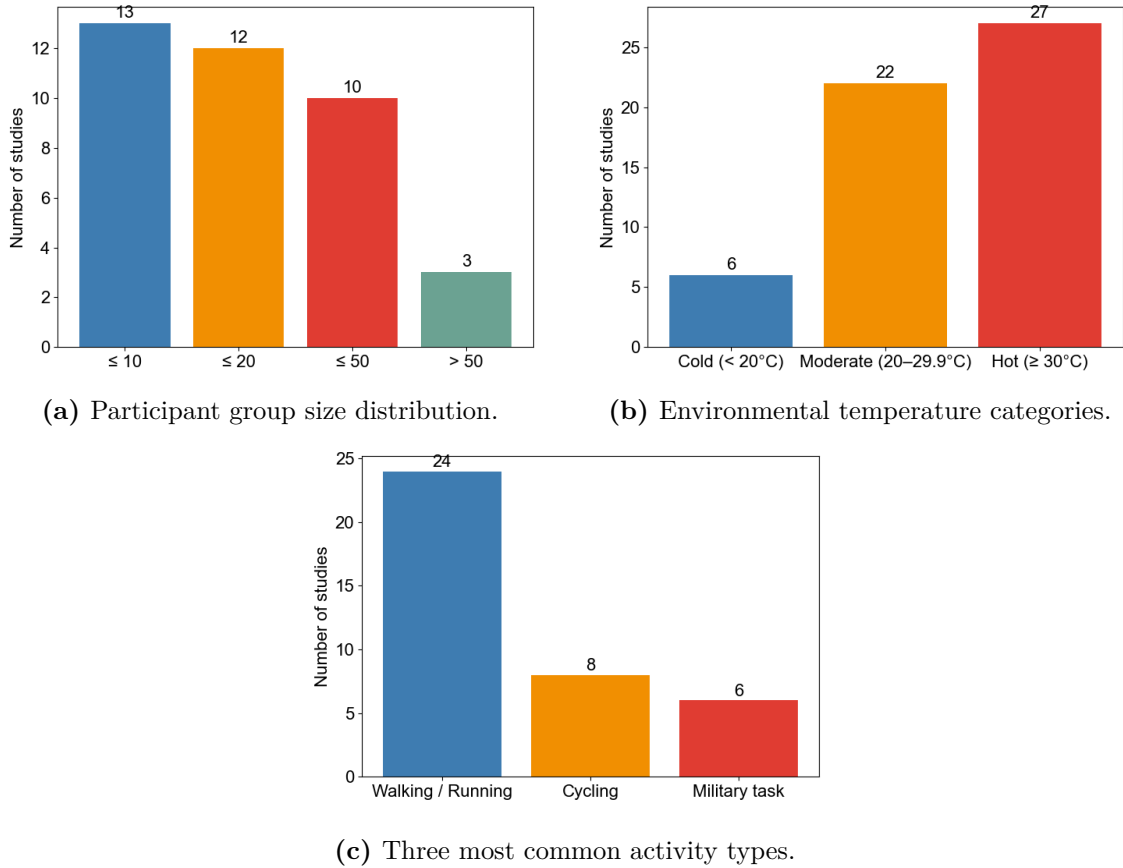


Figure 3.4: Overview of study characteristics across the included studies. (a) Participant group size distribution (all 38 studies). (b) Environmental exposure categories (Cold $<20^{\circ}\text{C}$, Moderate $20\text{--}29.9^{\circ}\text{C}$, Hot $\geq 30^{\circ}\text{C}$). One study did not report ambient temperature and is therefore not represented in panel (b). (c) Distribution of the three most frequently reported activity types (walking/running, military tasks, and cycling). Categories in (b) and (c) are not mutually exclusive, so a single study may contribute to multiple categories.

3.3.3 Reported performance across studies

Across the 38 included studies, the performance metrics reported in the original publications showed substantial heterogeneity. For the purpose of this systematic review, only five commonly reported accuracy measures were extracted: root mean square error (RMSE), mean absolute error (MAE), mean difference (MD), standard deviation of error (SD), and the coefficient of determination (R^2). These were the only metrics that appeared with sufficient frequency across studies to permit consistent extraction.

However, the level of detail varied considerably. Several studies reported full sets of overall error metrics (e.g., RMSE, MAE, MD, and SD), whereas others provided only a subset (e.g., reporting RMSE alone or R^2 only). A number of studies also reported site-specific or condition-specific values apart from the overall metric, and some studies did not provide any numerical performance values suitable for extraction.

Among the models that reported RMSE, values ranged from as low as approximately 0.02 °C (e.g., sweat–electrolyte–based random forest models) to over 0.60 °C in population-level mixed-effects models evaluated under field or occupational conditions.

Kalman filter and extended Kalman filter (EKF) studies generally reported RMSE values in the range of 0.13–0.41 °C, depending on the environmental setting, activity protocol, and whether phase-specific models (e.g., activity vs. recovery) or additional inputs such as skin temperature.

Statistical models exhibited similarly wide variability. Multiple regression approaches, which typically incorporated combinations of skin temperatures, heat-flux measurements, or environmental parameters, demonstrated substantial performance differences across studies. Depending on the specific anatomical sites included, the model formulation (e.g., PCA-based feature reduction or nonlinear mixed-effects structures), the presence or absence of heat-flux inputs, and additional specialized variables (e.g., insulated skin temperatures or ECG-derived Poincaré features), reported RMSE values ranged from 0.02 °C to 0.67 °C. This broad spread reflects heterogeneity in measurement site selection, the number and type of included physiological signals, and the diversity of environmental conditions and activity protocols.

Machine-learning approaches were less widely represented but exhibited performance ranges comparable to those of the statistical models. Two studies employing random forest regression reported markedly different performance depending on the input composition. When sweat-based biomarkers were included together with heart rate, RMSE values as low as 0.02°C were observed, whereas models relying on aural or auricle temperature together with heart rate reported RMSE values around 0.27°C . The only deep-learning-based formulation—a long-term sequence forecasting model (Informer) combined with or without EKF-derived features—showed a strong dependence on the number of skin-temperature sites used as input. Reported RMSE values ranged from approximately 0.28°C for heart-rate-only models to 0.12°C when seven-site skin-temperature arrays were included. These findings indicate that ML model performance is highly sensitive to input dimensionality and physiological signal diversity.

Hybrid models, which combined physics-based thermophysiological formulations with data-driven components, demonstrated similarly broad performance variability. Models incorporating first-principles thermoregulation or energy-balance equations together with autoregressive residual correction achieved RMSE values of approximately $0.22\text{--}0.30^{\circ}\text{C}$ in laboratory and field evaluations. Energy-balance formulations combined with adaptive Kalman filtering showed comparable performance, with reported RMSE values around 0.33°C . Others that integrated two-node thermoregulation models with parameter filtering reported small mean differences (e.g., -0.07°C) but did not provide full accuracy metrics.

3.3.4 Benchmark performance across the eligible studies

Figure 3.5 summarizes the benchmarked continuous CBT tracking accuracy of all 14 selected studies on Dataset 1 and Dataset 2, quantified by RMSE. For each model, detailed performance is provided in Table 3.4.

Across studies adopting the HR-only extended Kalman filter (EKF) originally proposed by Buller et al. [46], multiple publications reported RMSE values under different experimental conditions. Buller et al. [46] reported RMSE of $0.30 \pm 0.13^{\circ}\text{C}$, while subsequent evaluations reported RMSE of $0.21 \pm 0.11^{\circ}\text{C}$ [169], 0.32°C [170], $0.37 \pm 0.13^{\circ}\text{C}$ in hot conditions and $0.49 \pm 0.26^{\circ}\text{C}$ in warm conditions [52], 0.41°C [140],

and 0.37 ± 0.15 °C [141]. Under the standardized benchmark framework, the same HR-only EKF formulation yielded RMSE of 0.41 °C on Dataset 1 and 0.66 °C on Dataset 2.

For EKF models incorporating heart rate and skin temperature observations, the nonlinear EKF proposed by Seng et al. [137], which uses chest skin temperature and HR as inputs, achieved benchmark RMSE of 0.33 °C on Dataset 1 and 0.52 °C on Dataset 2. In the original study, Seng et al. [137] reported an RMSE of 0.29 °C.

Phase-dependent EKF variants showed slight improved benchmark RMSE values compared with HR-only EKF approaches. The modified EKF reported by Rizvi [138] achieved RMSE of 0.40 °C on Dataset 1 and 0.54 °C on Dataset 2, compared with a reported RMSE of 0.16 °C. The biphasic Kalman filter model proposed by Falcone et al. [148] yielded benchmark RMSE of 0.37 °C on Dataset 1 and 0.49 °C on Dataset 2, whereas the original publication reported RMSE of 0.28 ± 0.12 °C.

Linear Kalman filter models combining heart rate with skin temperature inputs demonstrated site-dependent benchmark performance. The linear KF reported by Buller et al. [172] achieved RMSE of 0.41 °C on Dataset 1 and 0.47 °C on Dataset 2, compared with a reported RMSE of 0.30 °C. For the linear KF proposed by Welles et al. [173], benchmark RMSE on Dataset 1 ranged from 0.36 °C for the thigh-based configuration to 0.45 °C for the forehead-based configuration, while on Dataset 2 the corresponding RMSE values were 0.68 °C and 0.53 °C. In the original study, reported overall RMSE values ranged from 0.28 to 0.30 °C depending on anatomical site.

Single-site statistical regression models based on skin temperature showed higher benchmark RMSE values. The linear regression models reported by Kim [174] yielded Dataset 1 RMSE values ranging from 0.42 °C to 0.57 °C across anatomical locations, with Dataset 2 RMSE values ranging from 0.51 °C to 0.55 °C. The original study reported site-specific R^2 values but did not report RMSE or MAE.

For the hybrid deep-learning EKF-Informer model proposed by Han et al. [175], the benchmarked heart-rate-only variant achieved RMSE of 0.41 °C on Dataset 1 and 0.34 °C on Dataset 2. In the original study, the reported RMSE for the heart-rate-only variant was 0.28 °C, with lower reported RMSE values observed when additional multi-site skin temperature inputs were included.



Figure 3.5: Benchmark comparison of continuous core body temperature tracking accuracy of selected studies, quantified by RMSE.

Figure 3.6 summarizes sensitivity and specificity for detecting elevated CBT using a threshold of 38.0°C on Dataset 1 and Dataset 2.

On Dataset 1, heart-rate-only EKF implementations yielded sensitivity values ranging from 0.71 to 0.75, with corresponding specificity values between 0.88 and 0.92. EKF models incorporating skin temperature measurements demonstrated higher sensitivity in most configurations, with values exceeding 0.77 for chest-based and thigh-based observations, while maintaining specificity above 0.85. Phase-dependent EKF variants, including those reported by Rizvi and Falcone, achieved sensitivity values between 0.73 and 0.74, with specificity consistently above 0.90.

Linear Kalman filter models combining heart rate with skin temperature exhibited

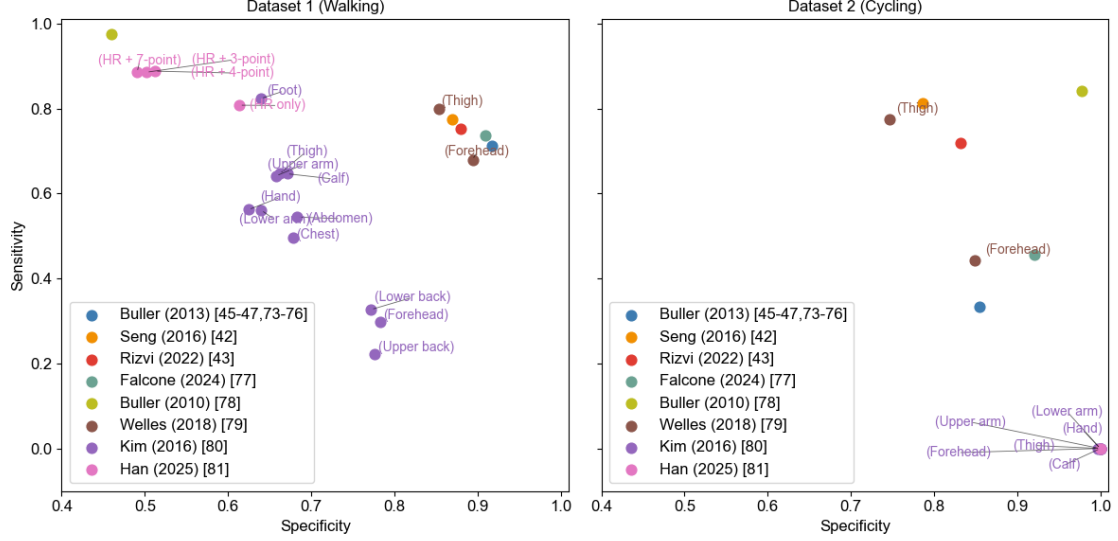


Figure 3.6: Sensitivity and specificity for high-risk core temperature detection (≥ 38.0 °C) across selected studies.

site-dependent classification performance. Thigh-based configurations achieved sensitivity values close to 0.80, whereas forehead-based configurations showed sensitivity below 0.70, despite maintaining specificity above 0.85. Single-site statistical regression models based solely on skin temperature demonstrated wide variability in sensitivity across anatomical locations, ranging from below 0.30 for upper-back measurements to above 0.55 for abdominal and chest locations, with specificity values generally exceeding 0.75.

On Dataset 2, sensitivity values were lower overall across most methods, while specificity remained comparatively high. Heart-rate-only EKF models yielded sensitivity values between 0.33 and 0.72, with specificity values between 0.83 and 0.85. EKF models incorporating skin temperature exhibited broader sensitivity ranges, from 0.45 to 0.81, depending on the observation configuration, with specificity values ranging from 0.78 to 0.92. Linear Kalman filter and regression-based models again showed strong site dependence, with some configurations achieving sensitivity above 0.70 while others remained below 0.40.

3.4 Discussion

This study adopted a two-tier analytical approach: first, a systematic review of data-driven methods for core body temperature (CBT) estimation, and second,

a standardized benchmark evaluation of the 14 algorithms that were able to re-implement using common wearable-derived inputs.

This systematic review of 38 studies provides the most comprehensive synthesis of modeling approaches used for non-invasive estimation of core body temperature during exercise, occupational heat exposure, or environmental stress. The included studies encompass a wide range of participant characteristics, activity protocols, environmental conditions, sensor configurations, and modeling frameworks, resulting in substantial heterogeneity in both methodology and reported performance. The present findings highlight clear trends in the evolution of core-temperature modeling, identify major sources of variability across the literature, and motivate the need for a standardized assessment framework.

Four broad modeling categories were identified across the included studies: Kalman filter or extended Kalman filter (EKF/KF) approaches, statistical models, machine-learning models, and hybrid models integrating physiological models with data-driven components. Statistical models were the most prevalent (17 studies), although EKF/KF approaches were also widely represented (13 studies), particularly in military and occupational settings. Pure machine-learning models were rare, appearing in only three studies, whereas hybrid formulations appeared in five studies.

Across the literature, heart rate was the most common input variable, followed by skin temperature, environmental conditions, load- or activity-related variables, and core temperature history. The diversity of input combinations, including the use of different body temperature sites, insulated skin temperatures, heat-flux measurements, and other specialized variables, was a major contributor to the heterogeneity in reported performance. Similarly, the choice of reference standard varied, with gastrointestinal telemetry pills used in half of the studies and rectal thermometry used in most of the remainder.

Substantial variability was also observed in participant sample size, environmental exposure, and activity protocols. More than half of the studies recruited fewer than 17 participants, and only three included more than 50 participants. Environmental temperatures ranged from cold conditions ($<20^{\circ}\text{C}$) to hot conditions ($\geq 30^{\circ}\text{C}$), and activity protocols included treadmill walking or running, cycling, load carriage,

military patrols, firefighting simulations, and field operations. These factors, combined with uneven reporting of methodological detail, make cross-study comparison difficult and reinforce the need for standardization.

Reported accuracy metrics varied widely across studies. Nevertheless, across the included literature the vast majority of reported RMSE and MAE values were within the 0.5 °C, a commonly adopted acceptance criterion in previous studies of core temperature estimation [147, 152].

The performance of Kalman filter and EKF models depended heavily on the observation model (e.g., quadratic HR–core temperature mapping, cubic skin temperature formulations), the presence of phase-specific measurement functions, and the environmental or activity context under which the filter was evaluated.

Statistical models demonstrated the greatest structural diversity, ranging from simple single-site regressions to high-dimensional multivariate formulations incorporating skin temperature, heat flux, and environmental variables. The wide variability in their reported performance reflects differences in sensor placement, the number of included sites, and the complexity of the underlying thermal conditions. Moreover, several studies did not report explicit accuracy metrics for these models, further limiting comparability across the literature.

Machine learning approaches, though less common, exhibited a marked dependence on the physiological richness of their input features. Models integrating multimodal or specialized variables such as sweat-related biomarkers or multi-site skin temperature measurements tended to perform substantially better than those relying solely on a single-site skin temperature. Overall, these observations suggest that incorporating a broader range of physiologically relevant input information can improve model accuracy.

Hybrid models, which combined first-principles thermophysiological formulations with data-driven components, showed promise for capturing physiological dynamics while retaining flexibility. However, these approaches also varied widely in structure and parameterization, and the lack of consistent reporting limited deeper comparison.

Taken together, these findings highlight the need for harmonized input definitions, standardized reference measurements, and unified accuracy metrics to enable mean-

ingful cross-study comparison. Future research would benefit from shared benchmark datasets, transparent reporting of model structures and parameters, and evaluation protocols that reflect realistic deployment conditions.

The review also highlights several broader implications for model development and evaluation. A central challenge is the absence of a common reference framework for comparing methods across studies. The heterogeneity in environmental conditions, sensor configurations, and experimental protocols makes direct comparison difficult, and this issue is further compounded by inconsistent reporting of performance metrics. Several studies provide only partial accuracy measures (e.g., RMSE without MAE or Bland–Altman statistics), while several studies omit numerical performance values entirely. These limitations hinder a clear assessment of relative model performance. This is particularly evident in the case of the widely used Buller 2013 extended Kalman filter: multiple studies have implemented the same method, yet reported different accuracy performance depending on the setting, population, heat exposure, and input availability. When such variability arises for the same method, cross-study comparison becomes even more difficult for distinct algorithms evaluated under heterogeneous protocols. This underscores the rationale for a standardized benchmark framework in which candidate models are re-implemented and evaluated under identical data, input, and evaluation settings.

The input requirements of many published models further constrain their practical deployability and the feasibility of head-to-head comparison. Of the 38 studies included in the systematic review, only 14 could be incorporated into the standardized benchmark analysis because the remaining methods depended on inputs that were not available in the benchmark datasets, such as insulated skin temperatures, multi-site heat-flux measurements, microclimate temperatures, or sweat-derived biomarkers. While these specialized signals can improve model fidelity in tightly controlled laboratory settings, they are often difficult, expensive, or impractical to obtain in real-world environments. In contrast, methods that rely on widely available signals such as heart rate or a single skin-temperature site are naturally aligned with commercially available wearables and therefore have clear translational potential. In the included studies, the heart-rate-only EKF was adopted and evaluated in several independent

investigations. Its repeated use reflects not only its methodological simplicity but, more importantly, its practical significance: by relying solely on heart rate, the method remains feasible in real-world athletic, occupational, and field settings where additional sensors may be unavailable, unreliable, or impractical [178–180]. The popularity of this minimalist-input formulation thus underscores its relevance for scenarios in which low-burden and broadly deployable solutions are essential.

The review also highlights limitations in current data resources for core-temperature modeling. More than half of the included studies enrolled fewer than 20 participants, indicating that many algorithms have been developed and tested on relatively small cohorts. Although such studies can still yield valuable physiological insight, they increase the risk of overfitting and limit the generalizability of reported findings across populations, environments, and clothing or equipment conditions. Together with the restricted input compatibility described above, these observations emphasise the need for larger, more diverse, and better documented datasets, as well as robust data-sharing mechanisms to support independent validation and comparative evaluation.

The standardized benchmark enables direct comparison of reported models under identical inputs, preprocessing, and evaluation settings, and it highlights several systematic performance gaps between the original publications and the reproduced results. For the HR-only EKF method derived from Buller et al. [46], the benchmark RMSE increased from the commonly reported range of $\sim 0.21\text{--}0.41^\circ\text{C}$ across prior studies [46, 52, 140, 141, 169, 170] to 0.41°C on Dataset 1 and 0.66°C on Dataset 2 (Table 3.4). Similar inflation was also observed for phase-dependent EKF variants (Rizvi [138]: reported 0.16°C vs. benchmark $0.40/0.54^\circ\text{C}$; Falcone [148]: reported $0.28 \pm 0.12^\circ\text{C}$ vs. benchmark $0.37/0.50^\circ\text{C}$ for Dataset 1/Dataset 2). These gaps suggest that the performance reported in the literature largely reflects the specific conditions of the study (participant group, environment, activity protocol, and sensor characteristics), rather than just the model structure itself. In particular, most EKF formulations rely on the stability of the HR–CBT mapping; shifts in exercise modality, thermal exposure, or wearable HR noise can directly alter observation residual statistics and thus degrade filtering performance when applied to a different dataset.

Across models, a consistent dataset dependence was observed. For multiple EKF/KF methods and statistical methods, RMSE was markedly higher on Dataset 2 than on Dataset 1. This result suggests that Dataset 2 presents a more challenging estimation setting, potentially due to differences in participant characteristics and activity protocols.

The benchmark also clarifies the effect of incorporating skin-temperature observations. Compared with the HR-only EKF baseline, the Seng EKF (HR + chest skin temperature) achieved lower RMSE on Dataset 1 (0.33°C vs. 0.41°C) and improved high-risk detection sensitivity on Dataset 2 (0.81), demonstrating that direct thermal observations can improve both continuous tracking and threshold-based detection when the selected site provides informative coupling to core temperature. However, the Welles linear KF showed strong site dependence: the variant with thigh and HR had lower RMSE than the variant with forehead and HR in continuous tracking on Dataset 1 (0.36 vs. 0.45°C), whereas on Dataset 2 the results are contrast (0.68 vs. 0.55°C). This indicates that the effectiveness of skin sites is not the same across all datasets and may be affected by multiple factors such as different participant group, environment, activity protocol, and sensor characteristics, leading to dataset-dependent performance.

For high-risk CBT detection, Fig. 3.6 shows that specificity was generally maintained at a high level across methods, whereas sensitivity exhibited substantially greater variability and was often lower on Dataset 2. Because models were trained and evaluated independently on each dataset, this asymmetry reflects differences in dataset-specific characteristics rather than transfer effects. In particular, sensitivity degradation on Dataset 2 suggests that, within this dataset, model prediction errors during high-temperature periods were more likely to remain below the fixed detection threshold, thereby reducing true-positive detections while preserving true-negative performance. This effect is further illustrated by several configurations that yielded sensitivity values of 0.00 on Dataset 2 (Table 3.4), indicating that predicted CBT trajectories rarely exceeded the 38.0°C threshold despite reference measurements doing so. These results illustrate that threshold-based detection outcomes can be highly sensitive to small biases in continuous CBT estimates. In particular,

modest underestimation near the 38.0 °C threshold can lead to a marked reduction in sensitivity, even when overall continuous error metrics remain moderate.

Although the standardized benchmark does not support a single universally best-performing method, several consistent patterns emerge. Across both datasets, EKF-based approaches relying solely on heart rate generally provided stable and interpretable performance and served as a strong baseline, particularly when input availability was limited. Variants that incorporated additional structure, such as phase-dependent observation models or auxiliary skin-temperature measurements, achieved improved accuracy under certain conditions, but these gains were not consistent across datasets.

Methods that depended on specific skin-temperature sites or more complex input configurations exhibited greater variability, with performance strongly influenced by sensor placement, participant characteristics, and experimental context. Taken together, these findings suggest that no single modeling strategy dominates across all settings; rather, simpler heart-rate-based filters are better suited to low-burden, broadly deployable scenarios, whereas multimodal approaches may offer advantages in controlled settings where informative skin-temperature measurements can be reliably obtained [178–180].

The present work has several limitations. First, although the systematic review has identified 38 eligible studies, only 14 algorithms have been included in the benchmark analysis. This restriction reflects the requirement that all benchmarked models must be reproducible using a standardized set of non-invasive inputs. As a result, several otherwise promising methods that rely on heat-flux measurements, insulated skin temperatures, or other specialized variables have not been eligible for inclusion. Second, model performance has been summarized primarily using aggregate error metrics such as overall RMSE and MAE across all samples. This approach has ensured a consistent population level comparison but may obscure meaningful inter-individual differences that are critical for personalized applications. Participant level metrics have not been emphasized because trial durations have varied considerably between individuals, and simple averaging of participant wise RMSE would have risked giving disproportionate weight to shorter recordings. Third,

the benchmark datasets themselves remain relatively small in terms of participant numbers and total trial count. This limits the diversity of physiological responses and environmental exposures represented in the analysis and reduces the statistical power available to evaluate algorithmic robustness across different subgroups or conditions [181]. In addition both datasets have been collected under controlled laboratory conditions and have involved predominantly young, healthy adults. The comparative findings therefore require cautious interpretation when extrapolated to older populations, clinical cohorts, or field-based occupational environments, where thermoregulatory dynamics, wearable sensor performance, and environmental constraints may differ substantially from those captured in the present data.

Recent developments in wearable sensing ecosystems further support the translational potential of non-invasive core temperature estimation. Multi-channel devices integrating inertial measurement units, ECG or PPG-based heart-rate sensors, and increasingly dense skin-temperature arrays are now commonplace in elite sports, military training centres, and occupational monitoring platforms [178–180, 182]. The growing availability of these multi-sensor systems provides a practical hardware foundation on which more sophisticated core temperature estimation algorithms can be deployed, enabling continuous, high-resolution monitoring under both controlled and free-living conditions.

These technological trends open up a wide range of future application scenarios for core temperature estimation. Accurate, real-time core-temperature monitoring could enable automated cooling or ventilation control systems, real-time heat-risk alerts during training and competition, and personalized hydration or pacing strategies tailored to an individual’s thermophysiological state [18–20]. Such applications highlight the broader importance of reliable core temperature estimation, not only as a research problem but as a key enabler for safety, performance optimization, and heat-illness prevention across athletic, military, and occupational domains.

4

Core Body Temperature Estimation from Heart Rate via Multi-Model Kalman Filtering and Variance-Based Fusion

Building on the findings of Chapter 3, which showed that empirical statistical models are the most commonly reported approach for non-invasive CBT estimation and that KF/EKF methods form a widely used second category, this chapter extends the EKF-based estimation to better adapt to physiological variability during physical activity. The focus is on developing a lightweight, real-time CBT estimation framework using heart rate (HR) as the sole input.

A state-aware, multi-model Kalman filtering framework with variance-based fusion is introduced, in which multiple state-specific EKFs are run in parallel and combined according to posterior uncertainty. Two variants are considered: a label-informed approach and an unsupervised approach that captures latent physiological variability without state labels.

Evaluated on two independent datasets, the proposed framework achieves RMSEs of 0.38–0.41 °C in within-dataset evaluations, lower than the selected EKF baselines. In cross-dataset analyses, the label-informed variant demonstrates improved robustness, achieving an RMSE of 0.88 °C, compared with 1.56 °C for the unsupervised variant.

Overall, these results demonstrate that incorporating physiological state awareness into EKF can improve estimation accuracy while keeping low computational complexity, and provide a methodological basis for the probabilistic state inference method in Chapter 5.

4.1 Introduction

Core body temperature (CBT) is a vital physiological indicator of thermal strain, particularly during physical activity or exposure to heat [183, 184]. In high-risk environments such as firefighting, military operations, and endurance sports, elevated CBT is strongly associated with heat-related illnesses including heat exhaustion, heat stroke, and exertional heat illness (EHI) [185, 186]. These conditions can develop rapidly and unpredictably, often in the absence of early subjective symptoms. If left unrecognized, dangerously high CBT levels may result in impaired cognitive and physical performance, loss of consciousness, organ failure, or even death [187–189]. Consequently, continuous and accurate CBT monitoring is critical not only for safeguarding individual health, but also for enabling data-driven strategies to manage thermal risk, guide exertion levels, and prevent life-threatening outcomes in heat-stressed environments [190–192].

Traditional methods of CBT assessment—such as rectal or esophageal thermometry, and ingestible telemetric pills—offer high accuracy but are invasive, cumbersome, and impractical for routine or prolonged use in the field [20, 147, 193, 194]. More accessible alternatives like axillary (underarm) and oral temperature measurements have been shown to poorly correlate with true CBT during periods of physical exertion or thermal challenge. These methods tend to underestimate actual thermal strain, making them unreliable for monitoring dynamic changes in core temperature [36, 37, 195]. As a result, there is a pressing need to develop and deploy effective, non-invasive CBT monitoring technologies that are suitable for real-world applications in occupational, athletic, and military settings.

Over the years, two broad approaches have emerged to address the challenge of non-invasive CBT estimation. One relies on physiological modeling grounded in the biophysics of human thermal regulation. Classical thermoregulatory models such as those by Stolwijk [47] and Fiala [48] simulate human thermal responses by dividing the body into multiple anatomical segments and tissue layers. These models compute dynamic heat transfer via conduction, convection, radiation, evaporation, and metabolic heat production, integrating inputs such as environmental conditions, metabolic rate, skin blood flow, and sweat rate [196–198]. While physiologically

detailed and useful for predictive simulation, these models require extensive parameterization, including clothing insulation and individual anthropometry, and are computationally intensive. Their assumptions of average physiological responses limit adaptability to individual variability, which is crucial for real-time, personalized monitoring in uncontrolled environments [199, 200].

The second approach leverages wearable sensing technologies to infer CBT indirectly through measurable physiological signals. Wearable systems offer the potential for continuous, unobtrusive, and real-time CBT estimation, making them highly attractive for field applications [50, 147, 154, 201]. Among various biosignals, heart rate (HR) has gained particular attention as a surrogate input for CBT due to its strong association with thermoregulatory activity [202–204].

The physiological relationship between HR and CBT reflects both thermoregulatory and metabolic demands. Under passive heat stress, rising CBT triggers thermoregulatory responses—such as increased skin blood flow and sweating—requiring higher cardiac output and thus elevating HR. Furthermore, during physical exercise, HR increases rapidly to meet metabolic demand, and the resulting internal heat production leads to a progressive rise in CBT [205, 206]. This empirical positive correlation between HR and CBT has been leveraged in various CBT estimation models, some of which rely solely on HR, while others incorporate additional physiological signals, using techniques such as linear regression [49, 150], machine learning and neural networks [175, 207], and state-space approaches like Kalman filters [138, 173].

Among these, Kalman filters (KFs) are particularly well-suited for real-time CBT tracking due to their ability to integrate prior knowledge of system dynamics with noisy observations. A typical KF model consists of a state transition function, which models the temporal evolution of CBT, and an observation function, which relates the observed HR to the hidden CBT state [46, 208]. The recursive structure of the KF enables real-time updates as new measurements arrive, making it attractive for wearable implementation.

Buller et al. [46, 169] proposed one of the earliest KF-based approaches for CBT estimation using HR. Their method uses an extended Kalman filter (EKF) with

a fixed quadratic observation model and a first-order autoregressive state model, assuming uniform physiological dynamics across individuals and activity phases. More recently, Falcone et al. [148] introduced the Biphasic Kalman Filter–Based (BKFB) model, which modifies Buller’s approach by dividing the estimation into two distinct phases—heating and cooling—based on predefined HR thresholds. Each phase uses a separate model to account for differences in thermoregulatory behaviour during CBT increase and decrease.

While both methods enable real-time, non-invasive CBT estimation, they rely on two key assumptions: (1) a universal HR–CBT mapping that ignores individual variability, and (2) clearly separable physiological states identifiable through simple HR thresholds. However, in real-world conditions, different states (e.g., rest and recovery) may yield similar HR patterns, and individuals exhibit substantial variability in thermal response. These limitations reduce model flexibility and generalizability.

To address these issues, a novel multi-model Kalman filtering framework with variance-based fusion for CBT estimation is proposed. In this framework, multiple independently parameterized Kalman filters are executed in parallel, and their outputs are dynamically fused based on posterior variance to produce a single, robust CBT estimate. Two complementary implementations of this framework are introduced:

1. Physiological State-Specific Kalman Filters (PSSK) leverage activity labels (e.g., *Rest*, *Exercise*, *Recovery*) to train separate models for each physiological state. This method assumes that activity annotations are available or can be reliably inferred.
2. Trial Clustering-Based Kalman Filters (TCBK) operate in a fully unsupervised manner by treating each participant-condition trial as an independent unit. A set of summary physiological features is extracted from each trial, including the mean, minimum, and maximum values of HR and CBT, as well as the slope of their linear relationship. K-means clustering is applied to group trials by latent physiological patterns, and cluster-specific Kalman filters are trained accordingly.

Both variants share a common fusion strategy but differ in how model specialization is achieved. While PSSK benefits from label-aware modeling when activity segmentation is available, TCBK provides a label-free alternative that can adapt to both individual and contextual variability. Both models are evaluated on two independent datasets, and their performance is compared against baseline approaches under within-dataset and cross-dataset conditions.

4.2 Method

4.2.1 Datasets

This study employed datasets from two previous studies involving healthy adult participants exposed to controlled thermal and physical activity conditions.

Dataset 1 includes 18 participants (10 males and 8 females) aged between 19 and 36 years [150, 151], which is a subset of the Dataset 1 described in Chapter 3. All sessions were conducted indoors. Each participant completed two experimental sessions corresponding to different clothing conditions: one with permeable cotton coverall and one with an impermeable coated nylon coverall. In each session, participants followed a structured protocol that began with 10 minutes of seated rest, followed by 40 minutes of treadmill walking. Afterward, participants rested until their rectal temperature had decreased by at least 0.4°C from the peak value. The session ended with the second 40-minute treadmill walking period. Throughout the experiment, heart rate, rectal temperature (as the ground truth for core temperature), and activity codes indicating the participant’s physiological state (e.g., rest or exercise) were continuously recorded.

Dataset 2 includes 13 male participants aged between 18 and 45 years [152]. Each participant completed two experimental sessions (heat stages 1 and 2), designed to simulate thermal strain under different workload and clothing conditions. In both stages, participants first rested seated outdoors for 15 minutes, then rested indoors for an additional 15 minutes. This was followed by a cycling session lasting between 20 and 60 minutes indoors. The exercise was terminated if any of the following conditions were met: rectal temperature increased by at least 1.5°C from baseline, rectal temperature exceeded 38.5°C , voluntary fatigue was reached, or a maximum of 75 minutes had elapsed. Finally, participants rested seated outdoors for 30 minutes.

In heat stage 1, participants exercised at 75% of their maximal heart rate ($HR_{\max}$) wearing light athletic clothing (T-shirt and shorts). In heat stage 2, the intensity was reduced to 50% of $HR_{\max}$, and participants wore full firefighter protective gear. Heart rate and rectal temperature were continuously recorded throughout each stage. No activity labels were provided in this dataset.

The demographic characteristics of participants in both datasets including average age, body mass, height, and body fat percentage are summarized in Table 4.1, along with details of the environmental temperature and humidity conditions, activity structure, and physiological variables used in this study.

Table 4.1: Summary of datasets. M=male, F=Female and HR=Heart rate

Attribute	Dataset 1 [150, 151]	Dataset 2 [152]
Number of participants	18 (10M / 8F)	13 (13M)
Age range (years)	19–36	18–45
Mean age ($\pm$ SD) in years	25.1($\pm$ 5.5)	30.9($\pm$ 5.4)
Mean body mass ($\pm$ SD) in kg	69.7($\pm$ 11.5)	77.5($\pm$ 6.1)
Mean height ($\pm$ SD) in cm	174.9($\pm$ 6.5)	179.2($\pm$ 6.4)
Mean body fat ($\pm$ SD) in %	14.9($\pm$ 5.2)	13.1($\pm$ 4.3)
Activity protocol & Environmental temperature and humidity	10 min rest(40°C, 20 %), 40 min walk(40°C, 20 %), rest until $\Delta T \geq 0.4^\circ\text{C}$ (22°C, 50 %), 40 min walk(40°C, 20 %)	15 min rest(23°C, 25 %), 15 min rest(35°C, 57 %), 20–60 min cycling(35°C, 57 %), 30 min rest(23°C, 25 %)
Stage conditions	permeable clothing(Setting 2), impermeable clothing(Setting 3)	75% of $HR_{\max}$, light clothing(Condition 1), 50% of $HR_{\max}$, firefighter gear (Condition 2)
Used variables	Heart rate (ECG chest based Equivital EQ-02+ LifeMonitor, Hidalgo, UK), rectal temperature (rectal probe, Edale Instruments, UK), activity code	Heart rate (ECG chest belt, Unico Swiss Tex, Switzerland), rectal temperature (rectal thermometer, MSR Electronics, Switzerland)
Sampling interval	1 minute	10 seconds

4.2.2 Multi-state Kalman Filters

To estimate core body temperature non-invasively from heart rate, a multi-model estimation framework built upon customized Kalman filters and a robust fusion strategy is designed. This section presents two complementary implementations: one based on predefined physiological states, and another relying on unsupervised participant-level clustering. Both approaches share a common structure: multiple

independently parameterized Kalman filters [46, 208] run in parallel, and their outputs are fused using a variance-based weighting mechanism to produce a final temperature estimate.

Each model k consists of a state transition model and an observation model. The state transition model describes how the hidden state evolves over time, whereas the observation model relates the hidden state to the available sensor measurements.

State Transition Model:

$$T_t^{(k)} = a_1^{(k)} T_{t-1}^{(k)} + a_0^{(k)} + \epsilon_t^{(k)}, \quad \epsilon_t^{(k)} \sim \mathcal{N}(0, \gamma_k^2) \quad (4.2.2.1)$$

The state transition model (Equation 4.2.2.1), also known as the process model, is formulated as a first-order autoregressive process [46, 177]. This reflects the physiological principle of thermal inertia, where the body's core temperature changes gradually, and its state at any given time is highly dependent on its immediate past. In this equation, $T_t^{(k)}$ denotes the estimated core temperature at time t under model k , and $T_{t-1}^{(k)}$ is the previous estimate. The coefficients $a_1^{(k)}$ and $a_0^{(k)}$ are learned from data and capture the autoregressive dynamics of temperature evolution. The process noise $\epsilon_t^{(k)}$ is assumed to follow a zero-mean Gaussian distribution with variance γ_k^2 .

Observation Model:

$$\hat{H}R_t^{(k)} = b_2^{(k)} \left(T_t^{(k)}\right)^2 + b_1^{(k)} T_t^{(k)} + b_0^{(k)} + \eta_t^{(k)}, \quad \eta_t^{(k)} \sim \mathcal{N}(0, \sigma_k^2) \quad (4.2.2.2)$$

The observation model (Equation 4.2.2.2) is based on the principle, established by Buller et al., that heart rate (HR) can serve as a powerful, albeit noisy, observation of the core temperature (CT). The physiological justification is that HR is a leading indicator of thermal strain, as it encapsulates critical information about both metabolic heat production and heat transfer [46]. Therefore, the model leverages this readily available signal to infer the hidden state of core temperature at each time step. In this equation, $\hat{H}R_t^{(k)}$ is the predicted heart rate based on the estimated temperature $T_t^{(k)}$ under model k . Coefficients $b_2^{(k)}$, $b_1^{(k)}$, and $b_0^{(k)}$ are learned separately for each model k . The observation noise $\eta_t^{(k)}$ is assumed Gaussian with zero mean and variance σ_k^2 .

At the beginning of the estimation process, each Kalman filter is initialized with a core body temperature of 37.0°C and an initial variance of 0.01 °C². Subsequently, each filter updates its core temperature estimate and associated uncertainty through a recursive process composed of prediction and update steps:

1. Prediction Step

$$\hat{T}_t^{(k)} = a_1^{(k)} T_{t-1}^{(k)} + a_0^{(k)} \quad (4.2.2.3)$$

$$\hat{v}_t^{(k)} = \left(a_1^{(k)}\right)^2 v_{t-1}^{(k)} + \gamma_k^2 \quad (4.2.2.4)$$

In this step, the prior estimate of the core temperature at time t under model k , denoted by $\hat{T}_t^{(k)}$, is predicted from the previous posterior estimate $T_{t-1}^{(k)}$ using the state transition model. The associated predicted variance $\hat{v}_t^{(k)}$ represents the uncertainty of the prior estimate, obtained by propagating the previous variance $v_{t-1}^{(k)}$ and adding the process noise variance γ_k^2 .

2. Observation Prediction

$$\hat{H}R_t^{(k)} = b_2^{(k)} (\hat{T}_t^{(k)})^2 + b_1^{(k)} \hat{T}_t^{(k)} + b_0^{(k)} \quad (4.2.2.5)$$

The expected heart rate $\hat{H}R_t^{(k)}$ is computed from the prior temperature estimate $\hat{T}_t^{(k)}$ using the observation model. The quadratic form reflects the physiological relationship between heart rate and core temperature under model k [46].

3. Extended Kalman filter mapping function variance coefficient

$$m_t^{(k)} = \frac{d\hat{H}R}{dT} = 2b_2^{(k)} \hat{T}_t^{(k)} + b_1^{(k)} \quad (4.2.2.6)$$

To enable Kalman-based updates with a nonlinear observation model, the function is linearized and the term $m_t^{(k)}$ denotes the extended Kalman filter mapping function variance coefficient.

4. Kalman Gain Computation

$$k_t^{(k)} = \frac{\hat{v}_t^{(k)} \cdot m_t^{(k)}}{(m_t^{(k)})^2 \cdot \hat{v}_t^{(k)} + \sigma_k^2} \quad (4.2.2.7)$$

The Kalman gain $k_t^{(k)}$ is a crucial parameter that is dynamically computed at each time step, where $k_t^{(k)}$ denotes the Kalman gain and the superscript (k) identifies the (k) -th parallel Kalman filter corresponding to a particular physiological state. Its primary role is to determine the weight given to the new observation (the heart rate measurement) versus the model's prediction when updating the state estimate. It is derived from the predicted state variance $\hat{v}_t^{(k)}$, the extended Kalman filter mapping function variance coefficient $m_t^{(k)}$, and the observation noise variance σ_k^2 . Specifically, a larger Kalman gain indicates that the observation is weighted more heavily than the model prediction, resulting in a stronger correction to the state estimate. Conversely, a smaller Kalman gain places greater weight on the model prediction, leading to a more conservative update. By continuously adjusting based on evolving uncertainty in both the model and measurements, the Kalman gain plays a central role in balancing model dynamics with real-time sensor data.

5. State Update

$$T_t^{(k)} = \hat{T}_t^{(k)} + k_t^{(k)} \cdot (HR_t - \hat{H}R_t^{(k)}) \quad (4.2.2.8)$$

The posterior estimate of core temperature $T_t^{(k)}$ is obtained by correcting the prior estimate $\hat{T}_t^{(k)}$ with the innovation term $HR_t - \hat{H}R_t^{(k)}$, which reflects the discrepancy between the actual and predicted heart rate. The correction is scaled by the Kalman gain.

6. Variance Update

$$v_t^{(k)} = (1 - k_t^{(k)} \cdot m_t^{(k)}) \cdot \hat{v}_t^{(k)} \quad (4.2.2.9)$$

The posterior variance $v_t^{(k)}$ is updated to reflect the reduced uncertainty after incorporating the current observation. The reduction factor $1 - k_t^{(k)} \cdot m_t^{(k)}$ indicates the information contribution of the observation relative to the prior uncertainty.

After each Kalman filter corresponding to model k independently completes the prediction and update steps, multiple parallel estimates $T_t^{(k)}$ are obtained, each associated with a posterior variance $v_t^{(k)}$. To produce a unified and robust estimate of core body temperature, these outputs are fused using a variance-based weighting strategy that reflects the confidence of each model.

$$w_k = \frac{1/v_t^{(k)}}{\sum_j 1/v_t^{(j)}} \quad (4.2.2.10)$$

$$T_t^{\text{fused}} = \sum_k w_k \cdot T_t^{(k)} \quad (4.2.2.11)$$

Here, w_k is the normalized fusion weight for model k , computed as the inverse of its posterior variance. This inverse-variance weighting scheme gives higher importance to models with greater confidence (i.e., lower uncertainty). The final fused estimate T_t^{fused} is thus a weighted average of all model-specific outputs, representing the best overall prediction under the assumption of independent Gaussian uncertainties.

Physiological State-Specific Kalman Filters (PSSK)

The dynamic relationship between heart rate and core body temperature varies depending on the underlying physiological state. Leveraging the activity labels available in Dataset 1, the data is categorized into three distinct states: *Rest*, *Exercise*, and *Recovery*. For each physiological state $k \in \{\text{Rest}, \text{Exercise}, \text{Recovery}\}$, a state transition model and a nonlinear observation model are independently estimated, as described in Section 4.2.2.

At inference time, three independent Kalman filters—each corresponding to a specific physiological state—are executed in parallel. All filters receive the same heart rate input but apply their own state-specific dynamics and observation models. Each filter outputs an updated estimate of core temperature along with its associated posterior variance. To obtain the final prediction, the outputs of all filters are fused using an inverse-variance weighting strategy, which reflects the relative confidence of each model.

Trial Clustering-Based Kalman Filters (TCBK)

To account for both individual variability and condition-dependent differences in thermoregulatory response, a data-driven framework called **Trial Clustering-Based Kalman Filters (TCBK)** is proposed. Unlike traditional participant-level clustering, TCBK treats each *trial*—defined as one participant completing a session under a specific condition—as an independent unit for clustering and model training.

For each trial, a set of summary features from the corresponding time series is computed, including the mean, minimum, and maximum values of heart rate and core temperature, as well as the slope of their linear relationship. These features are standardized via z-score normalization, and K-means clustering is applied to group trials into $K = 3$ clusters based on physiological similarity. To ensure optimal clustering quality, an exhaustive search over all possible combinations of two or more features is performed, and the subset that minimizes the downstream core temperature estimation root mean squared error (RMSE) is selected.

Once the trial clusters are formed, a separate Kalman filter for each cluster is trained. Each model shares the same structure as described in Section 4.2.2, with cluster-specific parameters for the state transition and observation models: $\{a_1^{(c)}, a_0^{(c)}, \gamma_c, b_2^{(c)}, b_1^{(c)}, b_0^{(c)}, \sigma_c\}$. These parameters are estimated using linear regression and second-order polynomial regression.

During inference, all cluster-specific Kalman filters operate in parallel, regardless of the original cluster assignment of the test trial. Each filter produces an independent estimate of core body temperature and its associated uncertainty. The final temperature estimate is then obtained via inverse-variance weighted fusion, where models with lower uncertainty contribute more heavily to the prediction.

This trial-level clustering approach provides greater flexibility and adaptability than participant-based clustering, as it captures variations across both individuals and experimental conditions. It is particularly advantageous in datasets where activity labels are unavailable or unreliable, and where physiological response patterns are highly condition-specific.

4.2.3 Model evaluation methods

To quantitatively assess the accuracy of core temperature estimation, this study employed three metrics: mean absolute error (MAE), root mean square error (RMSE), and Bland–Altman statistics. These metrics were computed by comparing estimated core temperature values against the ground truth at each time point [106, 209].

All metrics were calculated based on concatenated time-series data from all participants within each dataset, providing a comprehensive evaluation of group-level performance. This approach reflects the overall estimation accuracy across diverse individuals and conditions.

Additionally, Bland–Altman plots were used to visualize and analyze prediction residuals, offering insights into both systematic biases (mean difference) and the spread of errors (standard deviation). These plots also revealed temperature-dependent error patterns, making them particularly valuable for assessing model behaviour across the full physiological range, especially at elevated core temperatures.

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^N |\hat{T}_i - T_i| \quad (4.2.3.1)$$

Root Mean Square Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^N (\hat{T}_i - T_i)^2} \quad (4.2.3.2)$$

Mean Difference and Standard Deviation (MD, SD):

$$\text{MD} = \frac{1}{N} \sum_{i=1}^N (\hat{T}_i - T_i) \quad (4.2.3.3)$$

$$\text{SD} = \sqrt{\frac{1}{N} \sum_{i=1}^N [(\hat{T}_i - T_i) - \text{MD}]^2} \quad (4.2.3.4)$$

These statistics are used in Bland–Altman analysis to evaluate the agreement between the estimated and reference temperatures. MD reflects systematic bias, while SD indicates the spread of the residual errors.

All metrics are reported based on concatenated predictions across all participants in the same dataset, enabling group-level performance evaluation.

The algorithms were implemented and evaluated using Python 3.10, NumPy 2.0.2, SciPy 1.13.1, and scikit-learn 1.5.2. All analyses were conducted on a consumer laptop equipped with an AMD Ryzen 7 processor without GPU acceleration.

4.3 Results

4.3.1 Evaluation on Dataset 1

Figure 4.1 presents an overview scatter plot of Dataset 1, including heart rate, core temperature, and their associated activity states (Rest, Recovery, Exercise). The figure reveals that the physiological signals corresponding to different activity states exhibit overlap, indicating the difficulty of state discrimination based solely on heart rate and core temperature.

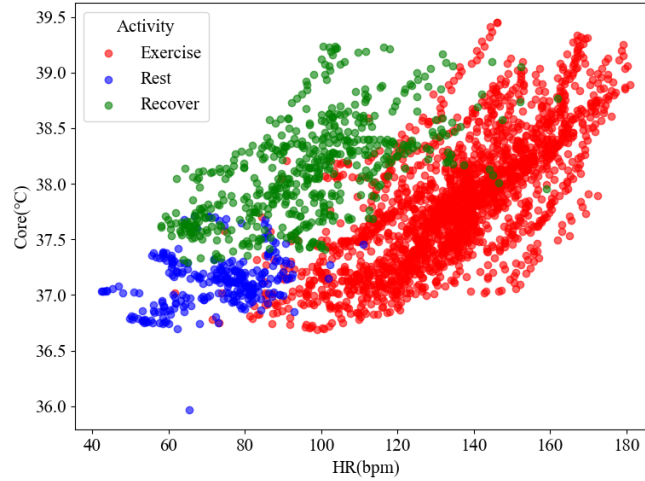


Figure 4.1: Heart rate, core temperature, and activity states in Dataset 1.

Table 4.2 summarizes the model parameters used in the four approaches evaluated on Dataset 1. For the PSSK model, separate Kalman filter parameters were trained for each physiological state. For TCBK, the optimal feature subset used for participant clustering is listed, along with the parameters for each resulting cluster-specific model. Buller’s model was fitted using a global parameter set, while the Falcone model includes separate parameter sets for the rising and falling phases, along with the best-performing heart rate threshold used for phase switching.

Table 4.2: Model parameters for PSSK, TCBK, Buller, and Falcone methods on Dataset 1

Model	State / Cluster	a_1	a_0	γ	b_2	b_1	b_0	σ
PSSK	Rest	1.0203	-0.7684	0.04732	-13.1008	988.3444	-18562.3877	11.6849
	Exercise	1.0049	-0.1602	0.04166	-6.2478	499.4499	-9816.3406	13.4916
	Recovery	1.0027	-0.1179	0.03080	-9.8319	776.9867	-15232.9121	15.3268
TCBK	Best Feature Set	<i>Max_Core, Min_Core</i>						
	Cluster 1	1.0038	-0.1273	0.0322	-8.9938	709.4774	-13844.8504	23.6935
	Cluster 2	0.9957	0.1727	0.0470	2.3500	-144.4650	2227.7564	24.7168
	Cluster 3	0.9663	1.2802	0.1119	-31.6569	2393.5826	-45118.5013	22.5549
Buller	—	1.0033	-0.1114	0.0447	-8.2635	652.4964	-12733.1383	24.2347
Falcone	HR Threshold	<i>120 bpm</i>						
	Rising Phase	0.9983	0.0928	0.0410	0.0142	13.7887	-401.3093	11.1216
	Falling Phase	0.9799	0.7481	0.0430	0.8260	-57.6493	1094.6583	16.9798

Figure 4.2 illustrates the estimated core body temperature over time using four different models on Dataset 1. The predictions from PSSK, TCBK, Buller’s model, and Falcone’s biphasic Kalman model are plotted alongside the ground truth values. The trials are concatenated along the time axis and visually separated by alternating grey and white background regions. Each trial is labelled using a three-digit identifier, where the first digit denotes the experimental condition and the final two digits correspond to the original participant number.

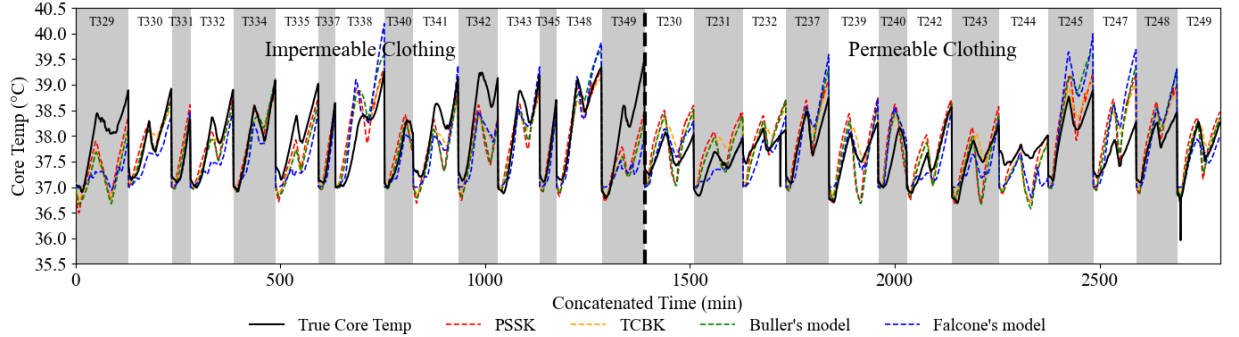


Figure 4.2: Prediction results on Dataset 1 using different models: The figure presents time-series plots of estimated core body temperature alongside the ground truth across all participants under two clothing conditions. Alternating grey and white background regions distinguish individual participants, enabling visual comparison of model performance across subjects. A vertical dashed line separates the two clothing conditions—impermeable on the left and permeable on the right. Each trial is labelled using a three-digit identifier, where the first digit denotes the experimental condition and the final two digits correspond to the original participant number.

Table 4.3 compares the performance of all four methods on Dataset 1 using RMSE, MAE, mean difference (MD), and standard deviation (SD) metrics. Both

PSSK and TCBK show lower RMSE than the baseline methods in all aspects.

Among the proposed methods, TCBK achieves the lowest RMSE (0.38°C), MAE (0.29°C), and SD (0.38°C), indicating its better accuracy and consistency. PSSK ranks second in RMSE (0.42°C), MAE (0.32°C), and SD (0.42°C), but achieves the best MD (0.01°C).

In contrast, Buller’s model shows moderate estimation accuracy, but with a negative bias ($\text{MD} = -0.06^{\circ}\text{C}$) and higher variance ($\text{SD} = 0.43^{\circ}\text{C}$). Falcone’s model performs the worst overall, with the highest RMSE (0.47°C) and the largest negative bias ($\text{MD} = -0.11^{\circ}\text{C}$), reflecting underestimation of core temperature.

Table 4.3: Performance comparison of CBT estimation methods on Dataset 1. The lowest obtained values are shown in bold.

Method	RMSE ($^{\circ}\text{C}$)	MAE ($^{\circ}\text{C}$)	MD ($^{\circ}\text{C}$)	SD ($^{\circ}\text{C}$)
PSSK	0.42	0.32	0.01	0.42
TCBK	0.38	0.29	-0.02	0.38
Buller et al.	0.43	0.33	-0.06	0.43
Falcone et al.	0.47	0.35	-0.11	0.46

Figure 4.3 presents the Bland–Altman plots for the four models, showing the agreement between estimated and ground truth core temperatures. PSSK and TCBK show tighter agreement and narrower limits of agreement compared to Buller’s and Falcone’s methods, especially at higher temperatures. The trials are concatenated along the time axis and visually separated by alternating grey and white background regions. Each trial is labelled using a three-digit identifier, where the first digit denotes the experimental condition and the final two digits correspond to the original participant number. Each trial is labelled using a three-digit identifier, where the first digit denotes the experimental condition and the final two digits correspond to the original participant number.

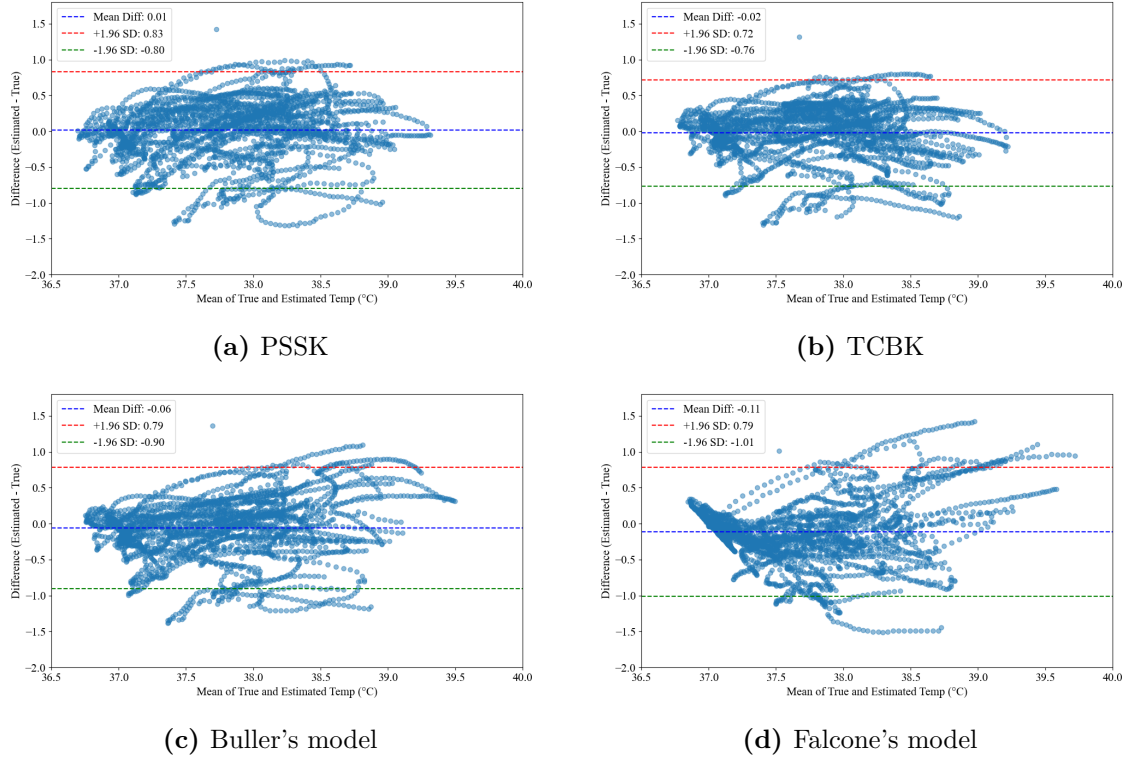


Figure 4.3: Bland–Altman plots of core temperature estimation on Dataset 1 using different models. The plots show the mean difference and 95% limits of agreement between predicted and ground truth temperatures.

4.3.2 Evaluation on Dataset 2

This section presents the evaluation results on Dataset 2, which consists exclusively of male participants under two thermal workload stages with distinct clothing conditions. Unlike Dataset 1, no activity labels were available, and the physiological states were not explicitly segmented. Thus, the PSSK model trained on Dataset 1 was directly transferred to Dataset 2 without retraining. For TCBK, Buller's, and Falcone's methods, new models were trained using Dataset 2. In total, seven methods were evaluated:

- Trial Clustering-Based Kalman Filters (TCBK, trained on Dataset 2)
- Buller's model (trained on Dataset 2)
- Falcone's model (trained on Dataset 2)
- PSSK (trained on Dataset 1)
- TCBK (trained on Dataset 1)

- Buller’s model (trained on Dataset 1)
- Falcone’s model (trained on Dataset 1)

Table 4.4 reports the model parameters for the methods trained on Dataset 2. The parameters for the transferred models are shown in Table 4.2 from Section 4.3.1.

Table 4.4: Model parameters for TCBK, Buller, and Falcone methods on Dataset 2

Model	State / Cluster	a_1	a_0	γ	b_2	b_1	b_0	σ
TCBK	Best Feature Set	<i>Avg_Core, Avg_HR, Min_Core, Max_HR</i>						
	Cluster 1	1.0007	-0.0231	0.0150	-34.4348	2594.7056	-48746.2651	34.6130
	Cluster 2	0.9993	0.0267	0.0162	-11.5513	880.5294	-16664.2499	27.1740
	Cluster 3	0.9987	0.0495	0.0153	-23.0177	1759.9794	-33519.5462	28.8622
Buller	—	0.9998	0.0086	0.0156	-10.6568	816.6215	-15521.8354	22.3432
Falcone	HR Threshold	<i>100 bpm</i>						
	Rising Phase	1.0000	0.0044	0.0154	-22.5097	1692.6964	-31678.6174	17.8573
	Falling Phase	0.9969	0.1145	0.0151	-0.5451	55.5554	-1234.3091	9.4995

Figure 4.4 illustrates the estimated core body temperature across all participants in Dataset 2 using the seven methods. Alternating grey and white background regions denote different participants, allowing visual comparison of model performance across individuals. The time-series plot demonstrates how well each model tracks the true temperature trend, highlighting differences in generalization performance for models transferred from Dataset 1 versus those trained on Dataset 2.

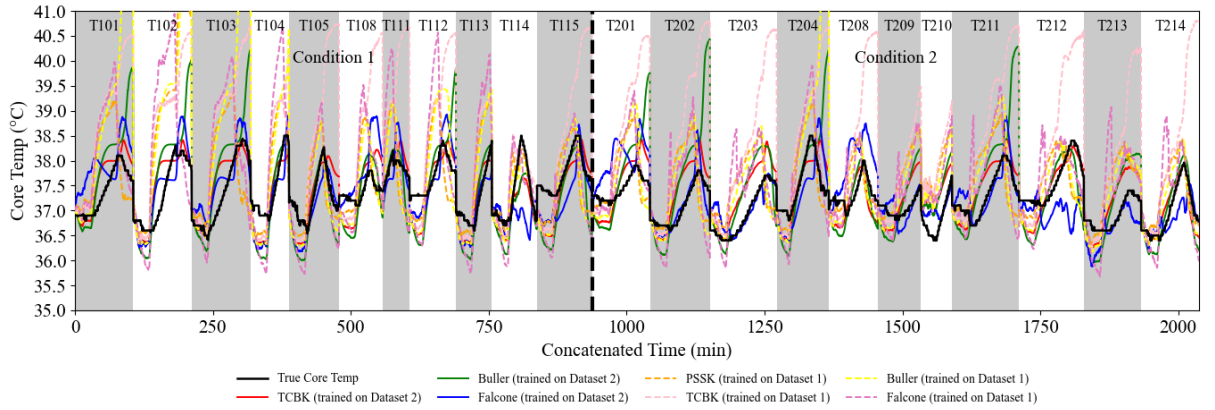


Figure 4.4: Prediction results on Dataset 2 using different models. The figure shows estimated and true core temperatures for all participants under different workload and clothing conditions.

Table 4.5 compares the estimation accuracy of all seven models on Dataset 2 using RMSE, MAE, mean difference (MD), and standard deviation (SD) of residuals.

Among the models trained directly on Dataset 2, the TCBK method achieves the best overall performance, with the lowest RMSE (0.41°C), MAE (0.35°C), and SD (0.41°C). Falcone’s model, while slightly inferior in RMSE and MAE, exhibits the lowest MD (0.01°C), indicating minimal systematic bias. Buller’s model performs worse than both TCBK and Falcone in all metrics, with a higher residual spread ($\text{SD} = 0.68^{\circ}\text{C}$) and moderate bias ($\text{MD} = 0.10^{\circ}\text{C}$).

In contrast, all models trained on Dataset 1 show significant performance degradation when applied to Dataset 2. The transferred TCBK model suffers the largest errors, with $\text{RMSE} = 1.56^{\circ}\text{C}$ and $\text{MAE} = 1.25^{\circ}\text{C}$, suggesting poor generalization of the learned clusters to unseen conditions. Buller’s and Falcone’s models trained on Dataset 1 also show elevated RMSEs (1.13°C and 0.97°C respectively), though slightly better than TCBK. PSSK performs the best among transferred models, achieving an RMSE of 0.88°C and MAE of 0.68°C , but still with large bias ($\text{MD} = 0.36^{\circ}\text{C}$) and variance ($\text{SD} = 0.80^{\circ}\text{C}$). This indicates that while PSSK retains some robustness, its performance still falls short compared to models retrained on Dataset 2.

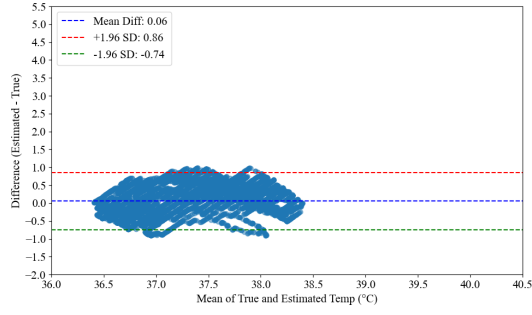
Table 4.5: Performance comparison of CBT estimation methods on Dataset 2. The lowest values are shown in bold.

Method	RMSE ($^{\circ}\text{C}$)	MAE ($^{\circ}\text{C}$)	MD ($^{\circ}\text{C}$)	SD ($^{\circ}\text{C}$)
TCBK (trained on Dataset 2)	0.41	0.35	0.06	0.41
Buller (trained on Dataset 2)	0.69	0.52	0.10	0.68
Falcone (trained on Dataset 2)	0.50	0.39	0.01	0.50
PSSK (trained on Dataset 1)	0.88	0.68	0.36	0.80
TCBK (trained on Dataset 1)	1.56	1.25	1.09	1.12
Buller (trained on Dataset 1)	1.13	0.79	0.43	1.04
Falcone (trained on Dataset 1)	0.97	0.74	0.33	0.91

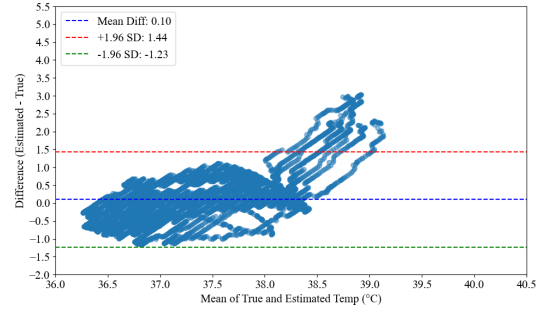
Figure 4.5 presents the Bland–Altman plots for all seven models. Among the models trained on Dataset 2, the TCBK method demonstrates the narrowest limits of agreement and the least variability across the full temperature range, indicating greater agreement across the range. Falcone’s model also shows relatively stable performance, though with slightly larger dispersion. In contrast, Buller’s model exhibits wider scatter and a more pronounced positive bias, particularly at higher core temperatures.

For the models transferred from Dataset 1, all exhibit larger limits of agreement and increased error variability. The transferred TCBK model shows the highest bias

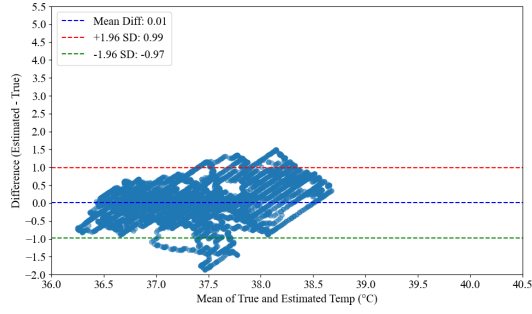
and variance, reflecting poor generalization under unseen thermal conditions. PSSK maintains closer alignment to the ground truth in terms of error trend, though it tends to overestimate core temperature at certain time points. Buller’s transferred model produces substantial overestimations in the high-temperature range. Although Falcone’s transferred model shows more uniformly distributed residuals, its overall error magnitude remains considerable.



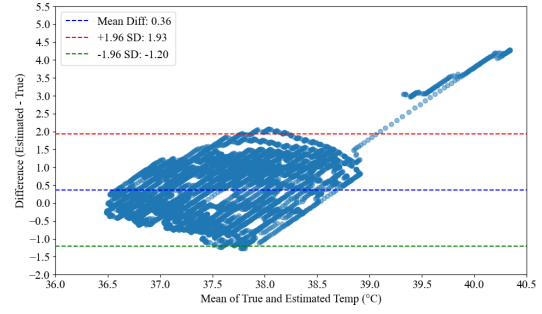
(a) TCBK (D2)



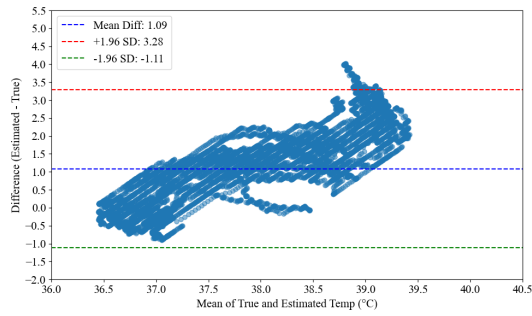
(b) Buller (D2)



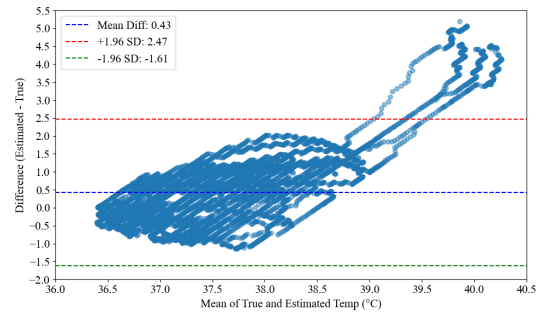
(c) Falcone (D2)



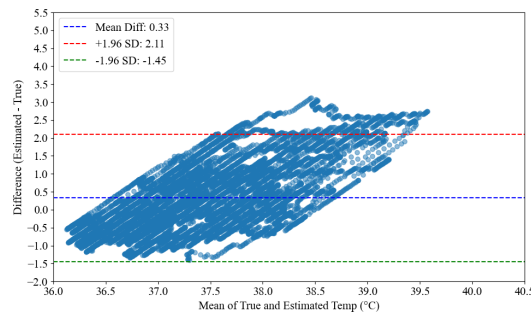
(d) PSSK (D1 to D2)



(e) TCBK (D1 to D2)



(f) Buller's model (D1 to D2)



(g) Falcone (D1 to D2)

Figure 4.5: Bland–Altman plots of core temperature estimation on Dataset 2 using different models. The plots show the agreement between predicted and ground truth values for models trained on Dataset 1 and Dataset 2.

4.4 Discussion

Two models for core body temperature (CBT) estimation were proposed: the Physiological State-Specific Kalman filter (PSSK) and the Trial Clustering-Based Kalman filter (TCBK). Both models are designed for implementation on wearable sensor devices. The PSSK model is trained using heart rate (HR), core temperature, and physiological state labels (e.g., rest, exercise, recovery), while the TCBK model is trained based on clustered trials that capture participant-specific and environmental variability.

During real-time CBT estimation, both models rely solely on heart rate as the input variable, which offers several advantages. Numerous commercially available wearable devices now provide reliable HR monitoring [210]. Moreover, relying exclusively on HR minimizes computational burden and sensor requirements, making the system lightweight and suitable for deployment on low-power wearable platforms. Compared to multimodal approaches that incorporate skin temperature or heat flux, HR-only models are easier to implement, and additional sensors may interfere with comfort and usability. It is also worth noting that skin temperature can vary considerably depending on sensor location and often deviates substantially from core temperature [35–37], which can reduce the consistency and generalizability of multimodal models.

However, as shown in Figure 4.1, the relationship between HR and CBT varies across different physiological states. The method proposed by Buller et al. assumes a static, one-to-one mapping between HR and CBT, which overlooks such variability and leads to substantial estimation error even within a single dataset. Falcone et al. attempted to address this by introducing phase-specific modeling, dividing the CBT trajectory into up-phase and down-phase based on HR thresholds. However, Figure 4.1 also demonstrates that there is considerable overlap in the HR–CBT relationship across different physiological states, indicating that such binary division is insufficient to capture the complexity of the underlying dynamics. As a result, Falcone’s model fails to generalize effectively and, in Dataset 1, even performs worse than Buller’s simpler model.

In contrast, the PSSK model explicitly captures state-dependent dynamics by training distinct Kalman filters for each physiological state and combining their predictions using confidence-weighted fusion. This approach accommodates both the diversity across states and similarities in overlapping regions. Consequently, PSSK consistently shows lower RMSE than baseline models in within-dataset and cross-dataset evaluations, demonstrating its capacity for better physiological representation.

To address the challenges of requiring real-time state labels, the TCBK approach was developed, which eliminates the need for explicit activity state annotations by clustering trials using descriptive features derived from HR and CBT signals. These include the mean, minimum, maximum, and slope of their linear relationship. This allows TCBK to model inter-individual and contextual variability, such as differences in baseline fitness or environmental stressors, without relying on direct physiological state labels. This contributes to TCBK’s strong performance in within-dataset evaluations.

Although both PSSK and TCBK demonstrate strong performance under controlled conditions, several limitations must be acknowledged. First, the relationship between HR and CBT is highly state-dependent. Even though HR is a useful proxy signal, its correlation with core temperature varies significantly across activity states, and any method that assumes a static HR–CBT relationship may suffer from systematic estimation bias, as evidenced by the distinct non-linear HR–CBT relationships across rest, exercise, and recovery shown in Figure 4-1. This issue is particularly important for HR-only models and necessitates careful modeling of context-specific dynamics, as addressed by PSSK.

Second, while PSSK effectively models such dynamics, it relies on accurate physiological state labels for training and deployment. In real-world applications, these labels are often unavailable or unreliable. Although the TCBK method avoids this requirement by clustering entire trials, it is more susceptible to overfitting to dataset-specific trends and shows weaker generalizability when applied to unseen datasets.

Third, both datasets used in this study feature participants with relatively homogeneous age distributions and were collected under limited environmental

variation. This may partially explain the high within-dataset performance of both models and raises concerns about their applicability to broader and more diverse populations.

Another practical limitation arises from the initialization requirement of Kalman filter-based methods. Accurate CBT estimation depends on the initial core temperature value at $t = 0$. However, in realistic deployment scenarios, such accurate initialization is often unavailable. In this study, the initial core temperature is set to 37.0°C with an initial variance of 0.01. Although this approximation is sufficient for experimental consistency, providing a more accurate initial CBT—e.g., via short-term calibration or one-time measurement—could further improve estimation performance in practical applications.

In addition, following the SAGER guidelines, whether model performance differed by sex was examined. Using Dataset 1, which includes male and female participants, the per-trial RMSEs for both models were compared. For PSSK, the aggregated RMSEs were 0.40 for males and 0.43 for females ($p = 0.71$, Mann–Whitney U test). For TCBK, the respective RMSEs were 0.32 and 0.43 ($p = 0.21$). In both cases, $p > 0.05$, indicating no statistically significant performance differences between sexes. These findings suggest that the proposed models generalize well across male and female participants in this dataset, although the small size of the dataset should be noted.

In summary, although multimodal approaches incorporating additional sensors may offer potential accuracy gains, the findings suggest that HR-only models strike a favorable balance between simplicity, performance, and deployability. Future work should explore adaptive fusion strategies that combine HR with other physiological signals in a context-aware manner, while also validating the proposed models across broader demographic and environmental settings.

4.5 Conclusion

Two Kalman filter-based models were proposed for core body temperature (CBT) estimation using only heart rate (HR): the Physiological State-Specific Kalman filter (PSSK) and the Trial Clustering-Based Kalman filter (TCBK). Both methods are lightweight and suitable for deployment on wearable devices.

PSSK incorporates physiological state labels to model state-dependent dynamics, while TCBK uses trial-level clustering based on HR and CBT statistics. Experiments on two datasets show that PSSK achieves robust performance across datasets, while TCBK performs best within datasets by capturing participant- and context-specific variability.

These results highlight the potential of HR-only models for accurate, efficient, and interpretable CBT estimation. Future work will explore adaptive state classification, multimodal integration, and large-scale validation in diverse populations and real-world environments.

5

Real-Time Estimation of Core Body Temperature for Heat Stress Monitoring in Hot Environments Using Wearable Heart Rate Sensors

The multi-model framework established in Chapter 4 demonstrated that adapting estimation strategies to different activity states can improve the accuracy of CBT estimation. By deploying multiple state-specific filters in parallel, this approach highlighted the importance of accounting for thermoregulatory dynamics across exercise and recovery/rest states. However, the variance-based fusion strategy adopted in Chapter 4 treats physiological states implicitly, relying on heuristic weighting rules rather than explicitly modeling how such states transition over time. This formulation assumes that thermoregulatory dynamics can be adequately approximated by a small number of distinct Gaussian state-space models, and does not explicitly represent uncertainty during transitions between activity states.

This chapter extends the method from heuristic fusion to probabilistic state inference. A model integrating a Hidden Markov Model (HMM) with a particle filter (PF) is proposed, in which physiological states (exercise and recovery) are treated as latent stochastic variables. A bidirectional coupling mechanism is introduced, whereby filter likelihoods are used to refine the HMM state belief, and the inferred state probabilistically modulates the contribution of state-dependent observation models within the filtering process. As a result, physiological state uncertainty is explicitly represented and propagated over time, rather than being handled indirectly through variance-based fusion weights.

Across two independently collected exercise datasets, the proposed probabilistic model achieved RMSEs of 0.359 °C and 0.371 °C, respectively, consistently lower than EKF-based baselines and their variants.

These findings show that explicitly accounting for uncertainty in physiological states leads to more accurate estimation when activity states are unclear. This chapter therefore demonstrates that physiologically informed probabilistic modeling offers a practical and effective alternative to rule-based approaches for monitoring heat strain.

5.1 Introduction

Despite the availability of preventive guidelines and environmental heat indices, the incidence of exertional heat illness remains persistently high [9, 211]. A key limitation of environment-based metrics is their inability to capture substantial inter-individual variability in thermoregulatory responses. This gap has motivated a growing interest in proactive, personalised physiological monitoring tools that can detect early signs of thermal strain. Among potential physiological markers, core body temperature (CBT) is particularly informative: it directly reflects deep tissue thermal status and is closely linked to the onset and severity of heat-related pathology. Continuous tracking of CBT therefore offers one of the most effective means of early detection, enabling timely interventions such as rest, cooling, hydration, or medical triage to prevent irreversible injury [212–214].

The clinical gold standard for measuring CBT is via invasive thermometry, with rectal and esophageal probes considered the most reliable reference methods [20, 38]. As recommended by the National Athletic Trainers’ Association (NATA), rectal temperature provides valid and stable measurements both at rest and during exercise, making it the criterion standard for assessing hyperthermic individuals in physically active settings [15]. However, the invasive nature, discomfort, and social barriers associated with these methods render them entirely unsuitable for routine, long-term, or ambulatory monitoring [31, 32]. Consequently, practitioners often resort to more convenient peripheral measurement sites, such as the axilla, oral cavity, or tympanic membrane [2, 215]. Unfortunately, these peripheral sites—such as the forehead, axilla, and oral cavity—not only deviate substantially from core temperature, but also lack

internal consistency with each other. Their readings are highly susceptible to ambient conditions and local blood flow, often producing conflicting or unpredictable values even under the same physiological state, particularly during physical activities [35–37]. Their demonstrated lack of accuracy under exertional heat stress makes them inadequate for applications requiring high-fidelity tracking of CBT, creating a distinct technological gap for a method that is both accurate and non-invasive.

To bridge this gap between the need for continuous monitoring and the limitations of invasive methods, research has historically pursued three major avenues: first-principles biophysical models, novel hardware-based sensors, and estimation models based on data from ubiquitous wearable devices. Classical thermoregulatory models, such as the seminal works by Stolwijk or Fiala, are grounded in first-principles heat balance equations [47, 48]. These multi-node models segment the human body into distinct anatomical layers (e.g., core, muscle, fat, skin) and simulate complex physiological processes like metabolic heat production, vasodilation, vasoconstriction, and evaporative cooling [196, 199]. While these models offer profound physiological interpretability and can predict temperature dynamics under various simulated conditions, their practical utility is severely hampered by a reliance on extensive and often inaccessible input parameters. Accurately measuring subject-specific metabolic rates, clothing insulation properties, ambient humidity, and convective air velocity in real-world scenarios is often infeasible. Furthermore, their deterministic structure can limit their ability to adapt to unmodeled physiological variability, reducing their robustness in uncontrolled environments [146, 148].

In parallel with modeling efforts, another avenue of research has focused on developing novel sensor technologies for direct, non-invasive CBT measurement. Ingestible telemetric capsules, for instance, are highly accurate and often used as a field standard, but they remain costly and impractical for long-term or widespread use [216, 217]. Technologies based on heat flux have also emerged as a promising alternative. The zero-heat-flux (ZHF) method uses a specialized probe, typically on the forehead, to create thermal equilibrium with deep tissue, showing good agreement with core temperature under certain conditions but remaining sensitive to probe placement [39, 40]. More recent wearable devices use heat flux (HF) sensors, often

in combination with other data streams, to estimate CBT [218]. While these sensors can achieve high accuracy, their performance can be affected by sweat and changes in skin blood flow during exercise [41]. The reliance of all these methods on specialized, dedicated hardware has sustained the drive to develop robust estimation models that can derive CBT from ubiquitous and accessible biosensors such as heart rate sensors [146, 147]. This approach is particularly valuable in scenarios requiring continuous and unobtrusive monitoring, including athletic performance tracking, heat acclimatization, and occupational safety.

These methods can be broadly categorized into static regression models and dynamic state-space frameworks. Static, data-driven approaches often employ supervised machine learning algorithms to learn a direct, nonlinear mapping from a vector of sensor inputs to a CBT value [49, 50, 150, 152, 167]. While often accurate under the conditions they were trained on, these multi-sensor models can be cumbersome, requiring extensive instrumentation, and may suffer from a lack of generalizability and interpretability, hindering their deployment in minimalist, real-world scenarios where only limited data streams are available.

Recognizing the practical limitations of multi-sensor inputs, a more structured approach has been to model the underlying physiological dynamics using a state-space framework. These models treat CBT as a latent state that evolves over time according to a process model and is indirectly observed via a measurement model [137, 139, 173]. This paradigm is particularly well-suited for single-input estimation from HR. The pioneering work in this domain was conducted by Buller et al., who first proposed an Extended Kalman Filter (EKF) to estimate CBT using HR as the sole observation input [46, 169]. This foundational model has since been widely adopted and validated across numerous conditions. Subsequent work by Buller’s group refined this model for periods of rest by replacing the original quadratic observation model with a generalized logistic function to better capture HR saturation behavior [219]. However, a key limitation of a single-model approach is its inability to account for the fundamentally different physiological regimes of active exercise versus post-exertional recovery [146].

To address this, Rizvi et al. introduced a modified EKF that employed a deterministic, rule-based "hard-switching" mechanism to alternate between two distinct observation models based on the trend of a moving average of HR [138]. While this represented a significant conceptual advance by incorporating state awareness, this deterministic switching can be brittle, struggling with ambiguity during gradual state transitions and exhibiting sensitivity to noise in the HR signal [220]. More fundamentally, this approach, along with all prior EKF-based models, is constrained by mathematical assumptions that do not fully capture physiological reality. The reliance on local linearization and a unimodal Gaussian assumption for the state distribution forces the model to assume a single, overly confident state at any given time [221]. This fails to represent the true uncertainty during complex physiological processes, such as the post-exercise recovery period where cardiovascular and thermoregulatory systems are dynamically recoupling. These simplistic modeling choices represent a persistent gap in achieving the robust and high-fidelity CBT estimation required for reliable physiological monitoring in the field.

To overcome these limitations, this paper proposes a comprehensive Bayesian filtering framework that enhances the robustness and applicability of core body temperature estimation from wearable sensors. First, a probabilistic state-fusion mechanism based on a Hidden Markov Model (HMM) bidirectionally coupled with Bayesian filters (EKF/PF) is introduced to replace brittle deterministic switching such as the BKFB in Chapter 4, thereby enabling reliable "soft-switching" between physiological regimes during exercise and recovery. An HMM represents the physiological regime, such as exercise or recovery, as a latent discrete state that cannot be observed directly. Instead, the probability of each state is inferred from the available physiological measurements and from transition probabilities describing how likely the system is to move from one regime to another over time [222]. This probabilistic representation enables soft switching, in which multiple physiological regimes may retain non-zero probabilities, rather than forcing an abrupt deterministic transition between models.

Second, an online adaptation strategy for the HMM is developed, allowing the model to continuously update its parameters and personalize to an individual's

time-varying dynamics. Third, a Particle Filter (PF) is employed to relax restrictive Gaussian and linearization assumptions, providing a non-parametric approach that better captures the nonlinear and non-Gaussian nature of human thermophysiology. Meanwhile, an Improved Genetic Algorithm (IGA) is integrated into the PF resampling process to improve particle diversity by generating new high-quality particles while preserving promising existing ones. Finally, nested cross-validation is adopted for hyperparameter optimization.

The proposed Online HMM-PF (IGA-Tuned) framework is evaluated against 9 alternative models, including reimplementations of foundational EKF approaches and ablation variants designed to isolate the contributions of the HMM, online adaptation, PF, and IGA components. Two datasets (Dataset 1 and Dataset 2) with diverse exercise and heat-stress protocols are used. Together, these advances establish a robust and practical framework that relies solely on heart rate sensing, making it directly compatible with widely available wearable devices. Leveraging this common biosensor signal, the approach directly estimates core body temperature in real time, providing a scalable pathway for physiological monitoring in sports, occupational safety, and healthcare.

5.2 Method

This section presents a unified Bayesian filtering framework for real-time CBT estimation using only HR signals. A common state-space model that captures the relationship between CBT and HR, including a mechanism to account for different physiological states, is first defined. The integration of this model within two filtering algorithms, the Extended Kalman Filter (EKF) and the Particle Filter (PF), is then described. For the PF, an advanced resampling strategy based on an Improved Genetic Algorithm (IGA) is introduced to enhance robustness. Finally, the comprehensive experimental protocol used to evaluate and compare all model variants on two distinct datasets is detailed.

5.2.1 Datasets

The methods are evaluated using two independently collected physiological datasets from prior experimental protocols involving healthy adult participants subjected to controlled thermal and physical activity exposures.

Dataset 1 consists of 22 healthy participants (12 males and 10 females) who completed treadmill walking protocols under six distinct environmental settings [150, 151]. These settings varied in ambient temperature (25–40°C), relative humidity (20–50%), solar radiation (0 or 530 W·m⁻²), and clothing type (permeable or impermeable). Each trial included alternating walking and rest phases, with temperature recovery targets. HR and rectal temperature were recorded at 1-minute intervals. Trials with missing HR or rectal temperature data were excluded, leaving 89 trials from all 22 participants for analysis. No participant was excluded entirely. Table 5.1 summarizes the demographic and protocol information.

Dataset 2 includes 13 male participants aged 18 to 45 years [152]. The same dataset was used as Dataset 2 in Chapter 4. The details of the participants, experimental protocol, and measurements are provided in the description of Dataset 2 in Table 4.1. HR was provided directly as an output from the chest ECG measurement systems.

For both datasets, separate 3-fold nested cross-validation was performed to train and evaluate all models.

5.2.2 The State-Space Framework for CBT Estimation

The work adopts and extends the state-space framework for CBT estimation originally proposed by Buller et al. [46, 169]. The task is formulated as inferring a latent state variable, $x_t \in \mathbb{R}$, representing the true core temperature at time t , from a sequence of noisy observations, $y_t \in \mathbb{R}$, representing the measured heart rate.

Process Model

Reflecting the body’s thermal inertia, the evolution of the core temperature state x_t is modeled as a first-order autoregressive process. This foundational model

Table 5.1: Summary of participant demographics and experimental settings of Dataset 1.

Attribute	Value
Number of participants	22 (12 males / 10 females)
Mean age ($\pm$ SD)	25.3 ± 5.2 years
Mean body mass ($\pm$ SD)	68.5 ± 10.9 kg
Mean height ($\pm$ SD)	173.8 ± 6.5 cm
Mean body fat ($\pm$ SD)	$15.7 \pm 6.2\%$
Clothing conditions	Permeable or impermeable coveralls
Environmental settings	• Setting 1: 25°C, 50% RH, impermeable, no solar • Setting 2: 40°C, 20% RH, permeable, no solar • Setting 3: 40°C, 20% RH, impermeable, no solar • Setting 4: 30°C, 35% RH, permeable, solar ($530 \text{ W}\cdot\text{m}^{-2}$ on back) • Setting 5: 40°C, 20% RH, permeable, solar ($530 \text{ W}\cdot\text{m}^{-2}$ on back) • Setting 6: 40°C, 35% RH, permeable, no solar
Activity protocol	• Setting 1: 10 min rest $\rightarrow$ 40 min treadmill walking $\rightarrow$ rest (22°C, 50% RH) until rectal temperature drops by 0.4°C $\rightarrow$ 40 min treadmill walking • Setting 6: 10 min rest in climatic chamber $\rightarrow$ 60 min treadmill walking $\rightarrow$ rest (22°C, 50% RH) until 0.4°C drop in rectal temperature $\rightarrow$ 40 min treadmill walking • Setting 2–5: Same protocol as Setting 1
Recorded signals	Heart rate (ECG chest based Equivital EQ-02+ LifeMonitor, Hidalgo, UK); rectal temperature (rectal probe, Edale Instruments, UK)
Sampling interval	1 minute

is shared across all evaluated methods:

$$x_t = a_1 x_{t-1} + a_0 + w_t, \quad w_t \sim \mathcal{N}(0, Q) \quad (5.2.2.1)$$

where a_1 and a_0 are scalar coefficients. For each fold in the cross-validation scheme, the parameters $\{a_1, a_0, Q\}$ are estimated from the corresponding training split via linear regression.

State-Dependent Observation Model

The relationship between the state x_t and the observation y_t is captured by a nonlinear function $h(x_t)$, which is fundamentally state-dependent. Two latent physiological states are defined, $s_t \in \{0, 1\}$, corresponding to a recovery phase ($s_t = 0$) and an exercise phase ($s_t = 1$). For each state, the observation model

is a separate second-order polynomial:

$$y_t = h_{s_t}(x_t) + v_t^{(s_t)}, \quad v_t^{(s_t)} \sim \mathcal{N}(0, R_{s_t}) \quad (5.2.2.2)$$

where the state-specific observation function is $h_s(x) = b_{2,s}x^2 + b_{1,s}x + b_{0,s}$. The coefficients $\{b_{k,s}\}$ and the state-dependent observation noise variance R_s are learned for each training fold as described below.

State-Specific Model Training

Within each training fold, the two observation models, h_0 for recovery and h_1 for exercise, are trained on data partitioned by a rule-based heuristic. Following the approach of Rizvi et al. [138], first, each data point in the current training split is labeled to create two distinct subsets, $\mathcal{D}_0$ and $\mathcal{D}_1$, for recovery and exercise, respectively:

1. A 5-minute moving average of the HR signal, $\bar{y}_t$, is computed to identify the underlying physiological trend.
2. If $\bar{y}_t < \bar{y}_{t-1}$, the time point is labeled as recovery. Otherwise, it is labeled as exercise.
3. The training data for the current fold is then split into $\mathcal{D}_0$ and $\mathcal{D}_1$ based on these labels.

Finally, the parameters for the h_0 model are fitted using only data from $\mathcal{D}_0$, while the parameters for h_1 are fitted using only data from $\mathcal{D}_1$. This process, repeated for each fold, yields specialized models that form the observation component for all subsequent state-aware filtering methods.

5.2.3 State Inference and Fusion Mechanisms

During the testing phase, the true physiological state s_t is unknown. Three distinct mechanisms for inferring this state were evaluated to select or combine the pre-trained observation models, h_0 and h_1 .

Rule-Based Switching

This is a deterministic, or "hard-switching," approach. The latent state is inferred at each time step using the same real-time HR trend rule described in Section 5.2.2. Based on the inferred state, the corresponding observation model (h_0 or h_1) is exclusively selected for the filtering update at that step. This mechanism is employed in the Switching EKF and Switching PF variants.

HMM-Based Fusion

A two-state Hidden Markov Model (HMM) provides a probabilistic, or "soft-switching," mechanism for combining the two state-specific observation models $h_s(\cdot)$ introduced in Section 5.2.2 [222]. The discrete physiological state takes values $s_t \in \{0: \text{recovery}, 1: \text{exercise}\}$. For each training fold, a transition matrix $\mathbf{A} \in \mathbb{R}^{2 \times 2}$ is estimated by empirical counting:

$$A_{ij} = \frac{\text{count}(s_{t-1} = i \rightarrow s_t = j)}{\text{count}(s_{t-1} = i)}. \quad (5.2.3.1)$$

During testing, the discrete-state belief is represented as a *row vector*

$$\mathbf{p}_t = \begin{bmatrix} p_t(0) & p_t(1) \end{bmatrix},$$

where $p_t(s) = P(s_t = s \mid y_{1:t})$. Following prior work, the initial belief is set deterministically to recovery: $\mathbf{p}_0 = [1, 0]$. At each time step, the HMM prediction step propagates the belief as

$$\mathbf{p}_{t|t-1} = \mathbf{p}_{t-1} \mathbf{A}, \quad (5.2.3.2)$$

which encodes the expected evolution of the physiological state before seeing the new HR sample.

1) State-conditioned emission likelihoods: Both EKF and PF based variants share the same state-dependent Gaussian observation model:

$$y_t \mid x_t, s_t = s \sim \mathcal{N}(h_s(x_t), R_s), \quad R_s = \sigma_s^2. \quad (5.2.3.3)$$

The marginal likelihood $p(y_t \mid s_t = s, y_{1:t-1})$ is evaluated using the current predictive distribution of x_t :

Particle filter (PF): For each particle $x_t^{(i)}$ with weight $w_{t-1}^{(i)}$,

$$p(y_t \mid s_t = s, y_{1:t-1}) \propto \sum_{i=1}^N w_{t-1}^{(i)} \exp\left(-\frac{1}{2}(y_t - h_s(x_t^{(i)}))^2 / R_s\right). \quad (5.2.3.4)$$

Extended Kalman filter (EKF): Given the one-step prediction $(x_{t|t-1}, P_{t|t-1})$,

$$p(y_t \mid s_t = s, y_{1:t-1}) = \mathcal{N}(y_t; h_s(x_{t|t-1}), H_s^2 P_{t|t-1} + R_s). \quad (5.2.3.5)$$

where $H_s = \partial h_s / \partial x$.

2) State update and filter coupling: The posterior HMM belief is updated by the forward recursion

$$p_t(s) \propto p(y_t \mid s_t = s, y_{1:t-1}) p_{t|t-1}(s), \quad (5.2.3.6)$$

followed by normalization. After normalization, the posterior $p_t(s)$ represents the final state weights at time t . These weights act as the soft-switching coefficients for fusing the two state-conditioned observation models in the subsequent continuous filtering step.

This posterior belief then influences the continuous filter:

- **PF**: particle weights use the fused likelihood

$$p(y_t \mid x_t^{(i)}) = \sum_{s \in \{0,1\}} p_t(s) \mathcal{N}(y_t; h_s(x_t^{(i)}), R_s). \quad (5.2.3.7)$$

- **EKF**: two parallel EKF updates (one for each s) are fused by moment matching using the weights $p_t(0)$ and $p_t(1)$.

Thus, the HMM and the continuous filter form an interacting hybrid system in which discrete-state inference and continuous CBT estimation inform each other at every time step. The architecture of this coupled system is illustrated in Figure 5.1 using the Particle Filter implementation.

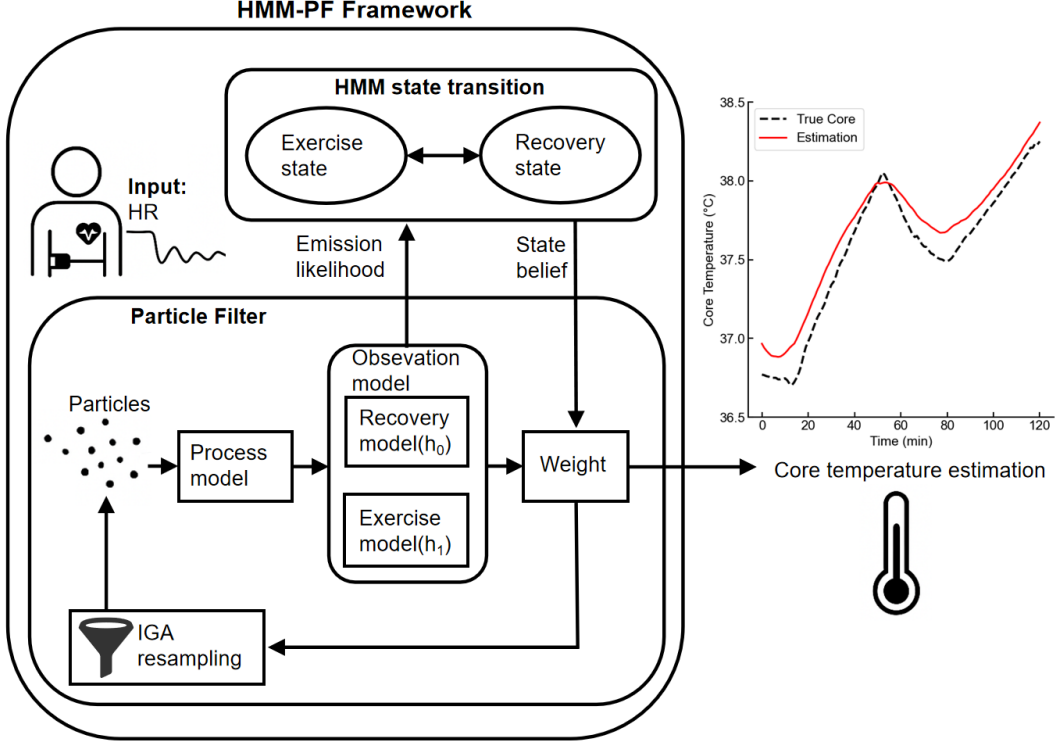


Figure 5.1: Schematic representation of the proposed coupled inference framework, illustrated here using the HMM-PF implementation. The diagram highlights the bidirectional information flow: the filter-derived emission likelihoods update the HMM state belief, which in turn provides soft-switching weights for the state-specific observation models (h_0 for recovery, h_1 for exercise). While the figure depicts particle-based estimation and IGA resampling, the HMM-EKF variant follows an analogous structure, replacing particle weighting with Gaussian moment matching.

Online HMM Adaptation

To accommodate non-stationary physiological behaviour, an online variant of the HMM is adopted, in which the transition matrix is updated using Dirichlet pseudo-counts. The expected transition probability is approximated as

$$P(s_{t-1} = i, s_t = j \mid y_{1:t}) \propto p_{t-1}(i) A_{ij} p(y_t \mid s_t = j, y_{1:t-1}), \quad (5.2.3.8)$$

and the Dirichlet parameters evolve according to

$$\alpha_{ij}^{(t)} = \lambda \alpha_{ij}^{(t-1)} + P(s_{t-1} = i, s_t = j \mid y_{1:t}), \quad (5.2.3.9)$$

with forgetting factor $\lambda = (M - 1)/M$ for a training-set memory of M samples. The adapted transition matrix is obtained by row-normalising $\alpha_{ij}^{(t)}$ and is used in the subsequent prediction step (5.2.3.2). This mechanism enables the Online-HMM EKF and Online-HMM PF to track gradual changes in the transition dynamics during testing.

5.2.4 Bayesian Filtering Algorithms

Two classes of Bayesian filtering algorithms are implemented and compared to solve the state estimation problem: the Extended Kalman Filter and the Particle Filter. Both leverage the state-space framework defined in Section 5.2.2 and the state inference mechanisms from Section 5.2.3.

Extended Kalman Filter (EKF)

The EKF implementation is based on the approach detailed by Buller et al. [46, 169]. The EKF assumes the state distribution is Gaussian, fully characterized by its mean $\hat{x}_t$ and variance P_t . As the state variable (CBT) is a scalar, the filter equations are implemented in their scalar form. The filter operates recursively in two steps:

1) Prediction Step The state estimate is propagated forward in time using the process model (Eq. 5.2.2.1):

$$\hat{x}_{t|t-1} = a_1 \hat{x}_{t-1|t-1} + a_0 \quad (5.2.4.1)$$

$$P_{t|t-1} = a_1^2 P_{t-1|t-1} + Q \quad (5.2.4.2)$$

where $\hat{x}_{t|t-1}$ and $P_{t|t-1}$ are the predicted state mean and variance, respectively.

2) Update Step The predicted state is corrected using the current measurement y_t . The EKF linearizes the observation model $h_s(x)$ using its Jacobian (which is a scalar derivative in this case), $H_{t,s} = \nabla_x h_s(x)|_{\hat{x}_{t|t-1}}$. The update proceeds as follows:

$$K_t = \frac{P_{t|t-1} H_{t,s}}{H_{t,s}^2 P_{t|t-1} + R_s} \quad (5.2.4.3)$$

$$\hat{x}_{t|t} = \hat{x}_{t|t-1} + K_t (y_t - h_s(\hat{x}_{t|t-1})) \quad (5.2.4.4)$$

$$P_{t|t} = (1 - K_t H_{t,s}) P_{t|t-1} \quad (5.2.4.5)$$

where K_t is the Kalman Gain.

The filter is initialized with a state estimate of $\hat{x}_0 = 37.0^\circ\text{C}$ and an initial variance of $P_0 = (0.2^\circ\text{C})^2 = 0.04$.

These steps are integrated with the state-aware mechanisms as follows:

- **Single EKF:** Employs one global observation model for all updates.

- **Switching EKF:** At each time step, it selects a single state $s_t \in \{0, 1\}$ and performs the EKF update using the corresponding model parameters based on the rules described in Section 5.2.2.
- **HMM-EKF / Online-HMM EKF:** These variants receive the soft state probabilities $\mathbf{p}_t = [p_t(0), p_t(1)]$ from the HMM (Section 5.2.3). Two parallel EKF updates are performed, one under each state hypothesis $s = 0, 1$, resulting in $(\hat{x}_{t|t}^{(s)}, P_{t|t}^{(s)})$. The final posterior is obtained by moment-matching the mixture of the two Gaussians weighted by $\mathbf{p}_t$:

$$\hat{x}_{t|t}^{\text{fused}} = \sum_{s \in \{0,1\}} p_t(s) \hat{x}_{t|t}^{(s)}, \quad (5.2.4.6)$$

$$P_{t|t}^{\text{fused}} = \sum_{s \in \{0,1\}} p_t(s) \left[P_{t|t}^{(s)} + \left(\hat{x}_{t|t}^{(s)} - \hat{x}_{t|t}^{\text{fused}} \right)^2 \right]. \quad (5.2.4.7)$$

The Online-HMM EKF uses the same formulation, but with an adaptively updated transition matrix.

Particle Filter (PF)

The PF is a non-parametric filter that approximates the posterior distribution with a set of N weighted samples or ‘particles’, $\{x_t^{(i)}, w_t^{(i)}\}$, making it suitable for non-Gaussian and highly nonlinear systems [223, 224].

The PF is initialized by drawing the N particles from a Gaussian distribution centered around a typical physiological value: $x_0^{(i)} \sim \mathcal{N}(37.0^\circ\text{C}, (0.2^\circ\text{C})^2)$. In the implementation, N is set to 200 to balance computational cost and estimation accuracy. This creates an initial particle cloud representing the prior belief of the state distribution. All particles are initially assigned equal weights, $w_0^{(i)} = 1/N$.

The algorithm then iterates through three steps:

1) Prediction Step Each particle is propagated independently through the stochastic process model (Eq. 5.2.2.1):

$$x_t^{(i)} = a_1 x_{t-1}^{(i)} + a_0 + w_t^{(i)}, \quad w_t^{(i)} \sim \mathcal{N}(0, Q) \quad (5.2.4.8)$$

2) Weight Update Step The weight of each particle, $w_t^{(i)}$, is updated based on the likelihood of observing the measurement y_t given the particle's state $x_t^{(i)}$:

$$w_t^{(i)} \propto w_{t-1}^{(i)} \cdot p(y_t | x_t^{(i)}) \quad (5.2.4.9)$$

After updating all particle weights, they are normalized to sum to one.

3) Resampling Step To prevent particle degeneracy (where few particles have all the weight), a resampling procedure is applied when the effective sample size ($N_{\text{eff}} = 1 / \sum_i (w_t^{(i)})^2$) drops below a threshold. While standard methods such as systematic resampling can be used, the framework primarily utilizes an advanced strategy, the Improved Genetic Algorithm (IGA), to enhance particle diversity. The IGA procedure is detailed in Algorithm 1.

These steps are integrated with the state-aware mechanisms as follows:

- **Single PF:** Employs one global observation model for all updates.
- **Switching PF:** At each time step, it selects a single state $s_t \in \{0, 1\}$ and performs the PF update using the corresponding model parameters based on the rules described in Section 5.2.2.
- **HMM-PF / Online-HMM PF:** These variants use the HMM-based state inference described in Section 5.2.3. At each time step, the HMM provides posterior state probabilities $\mathbf{p}_t = [p_t(0), p_t(1)]$, which act as soft-switching weights for the two observation models. The particle-wise likelihood is computed as a mixture over both state hypotheses:

$$p(y_t | x_t^{(i)}) = \sum_{s \in \{0, 1\}} p_t(s) \mathcal{N}(y_t; h_s(x_t^{(i)}), R_s). \quad (5.2.4.10)$$

The Online-HMM PF shares the same likelihood formulation, but updates the transition matrix online as detailed above.

After likelihood evaluation, all particle weights are normalised, and the final CBT estimate is obtained as the weighted mean $\hat{x}_t = \sum_i w_t^{(i)} x_t^{(i)}$.

Algorithm 1 Improved Genetic Algorithm (IGA) Resampling

Input: Particles $\{x_t^{(i)}\}_{i=1}^N$, Weights $\{w_t^{(i)}\}_{i=1}^N$

Hyperparameters: Crossover probability p_c , Mutation probability p_m , Mutation strength σ_m , Concentration threshold ϵ_{conc}

Output: New particles $\{\tilde{x}_t^{(i)}\}_{i=1}^N$

Step 1: Resampling Trigger

- Calculate effective sample size: $N_{\text{eff}} = \left(\sum_{i=1}^N (w_t^{(i)})^2\right)^{-1}$.
- If $N_{\text{eff}} \geq N/2$, no resampling is needed. Return original particles $\{x_t^{(i)}\}$ and terminate.

Step 2: Probabilistic Parent Selection

- *a) Estimate Concentration:* Compute local concentration for each particle:

$$c^{(i)} = \sum_{j=1}^N \mathbb{I}(|x_t^{(i)} - x_t^{(j)}| < \epsilon_{\text{conc}})$$

- *b) Adjust Probabilities:* Compute selection probabilities that penalize high-concentration particles:

$$P^{(i)} = \frac{w_t^{(i)} \cdot \exp(-c^{(i)})}{\sum_j w_t^{(j)} \cdot \exp(-c^{(j)})}$$

- *c) Select Parents:* Sample N parent particles $\{x_p^{(i)}\}$ from the original set $\{x_t^{(i)}\}$ according to probabilities $\{P^{(i)}\}$.

Step 3: New Particles Generation via Crossover and Mutation

- *a) Crossover:* Partition parents $\{x_p^{(i)}\}$ into pairs. With probability p_c , apply arithmetic crossover to each pair to generate an intermediate set of particles $\{x_c^{(i)}\}$.
- *b) Mutation:* For each intermediate particle $x_c^{(i)}$, apply mutation with probability p_m by adding Gaussian noise. This produces the final new particles:

$$\tilde{x}_t^{(i)} = x_c^{(i)} + \eta, \quad \text{where } \eta \sim \mathcal{N}(0, \sigma_m^2)$$

Step 4: Output

- Return the final set of new particles $\{\tilde{x}_t^{(i)}\}$. Their weights are reset to be uniform ($w_t^{(i)} = 1/N$).
-

5.2.5 Experimental Protocol and Evaluation

A rigorous protocol was designed to ensure a fair and robust comparison of all model variants.

Model Configurations and Evaluation Protocol

A suite of 10 models was evaluated, comprising the logical combinations of the described filters (EKF, PF), state-inference mechanisms (Single, Switching, HMM, Online-HMM), and resampling strategies (Standard, IGA). All models were trained and tested using a nested 3-fold cross-validation scheme stratified by experiment trials [225]. To account for stochasticity, the entire protocol was repeated 10 times with different random seeds. The final prediction for any given time point is the average of the predictions from all 30 relevant runs (3 folds $\times$ 10 seeds). The algorithms were implemented and evaluated using Python 3.10, NumPy 2.0.2, SciPy 1.13.1, and scikit-learn 1.5.2. All analyses were conducted on a consumer laptop equipped with an AMD Ryzen 7 processor without GPU acceleration.

Hyperparameter Tuning

For all IGA-based models, hyperparameters $(p_c, p_m, \sigma_m, \epsilon_{\text{conc}})$ were optimized within the inner loop of the nested cross-validation using the Optuna framework [226]. The optimization objective was to minimize the average RMSE over 100 optimization trials on the inner-loop validation sets.

Performance Metrics

Model performance was assessed on the held-out test sets using the following metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Determination (R^2), and the Mean Difference (MD) $\pm$ Standard Deviation (SD) of the estimation error for Bland-Altman analysis [106, 209].

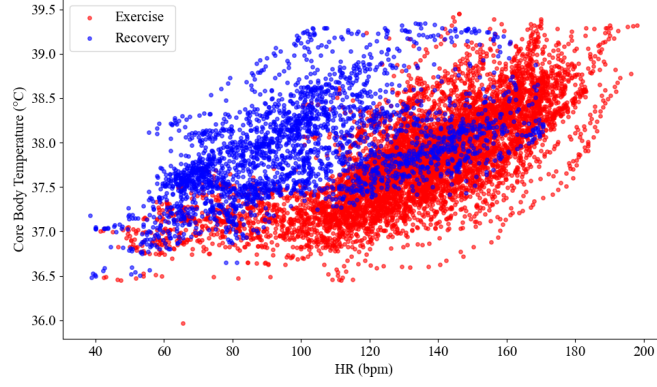
5.3 Results

This section presents both qualitative and quantitative evaluations of model performance across the two experimental datasets.

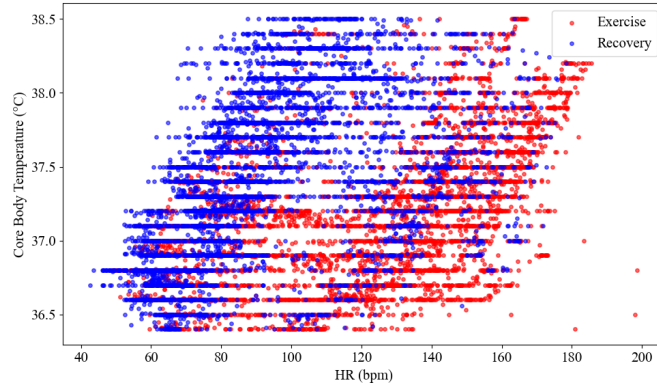
To better understand the underlying relationship between heart rate (HR) and core body temperature (CBT), the raw data distributions from both datasets are first visualized, as shown in Figure 5.2. Each point represents a sample colored by physiological state—either *exercise* or *recovery*—as determined by the rule-based labeling scheme described in Section 5.2.2. Specifically, the state at each time point is inferred from the trend of a 5-minute moving average of HR: decreasing trends indicate recovery, while non-decreasing trends indicate exercise.

Figure 5.2a corresponds to Dataset 1, and Figure 5.2b corresponds to Dataset 2. In both datasets, partially overlapping yet distinguishable HR–CBT patterns are observed across physiological states. While both *exercise* and *recovery* phases exhibit a general trend in which core body temperature (CBT) increases with heart rate (HR), their distributions differ markedly. The data points during exercise are more tightly clustered, indicating a more stable HR–CBT relationship, whereas those during recovery are more dispersed, reflecting greater physiological variability and transitional effects. Additionally, the overall trend during recovery tends to be shallower compared to exercise.

These distributional differences—both in spread and directional trends—highlight the necessity of modeling physiological states explicitly. To handle the overlapping regions and smooth transitions between states, the framework employs a soft fusion strategy: observation likelihoods from state-specific models are combined using weights derived from a hidden Markov model (HMM). This probabilistic fusion enables the filter to adapt continuously to ambiguous or transitional data points, improving robustness and estimation accuracy.



(a) Dataset 1



(b) Dataset 2

Figure 5.2: Scatter plots of heart rate (HR) versus core body temperature (CBT) for (a) Dataset 1 and (b) Dataset 2. Each point is colored by physiological state—*exercise* (red) or *recovery* (green)—as determined by the rule-based labeling scheme described in Section 5.2.2.

The comprehensive performance metrics for all models on Dataset 1 and Dataset 2 are presented in Table 5.2 and Table 5.3, respectively.

The Online HMM-PF (IGA) model yielded the lowest Root Mean Square Error (RMSE) on both datasets, recording a value of 0.359°C on Dataset 1 and 0.371°C on Dataset 2. The top-performing models in all cases were variants of the Particle Filter (PF), particularly those incorporating the Improved Genetic Algorithm (IGA) resampling. In contrast, the Extended Kalman Filter (EKF) variants consistently resulted in higher estimation errors. Notably, on Dataset 2, the Switching EKF and Single EKF models produced negative R^2 values of -0.126 and -0.549, respectively.

Figure 5.3 presents time-series comparisons between the estimated and reference core temperatures for selected test trials from Dataset 1 and all test trials from Dataset 2. Each subplot corresponds to one dataset, with trials concatenated

Table 5.2: Performance Comparison of All Models on Dataset 1.

Method	RMSE	MAE	R^2	MD	SD
Online HMM-PF (IGA)	0.359	0.276	0.571	+0.003	0.359
HMM-PF (IGA)	0.359	0.276	0.570	+0.003	0.359
Online HMM-PF (Standard)	0.366	0.281	0.554	+0.014	0.366
HMM-PF (Standard)	0.367	0.282	0.551	+0.015	0.367
Switching PF (Standard)	0.372	0.287	0.541	-0.035	0.370
Online HMM-EKF	0.380	0.290	0.519	-0.035	0.379
HMM-EKF	0.380	0.290	0.519	-0.035	0.379
Single PF (Standard)	0.383	0.291	0.512	-0.073	0.376
Switching EKF	0.395	0.305	0.481	-0.034	0.394
Single EKF	0.421	0.316	0.411	-0.074	0.414

Table 5.3: Performance Comparison of All Models on Dataset 2.

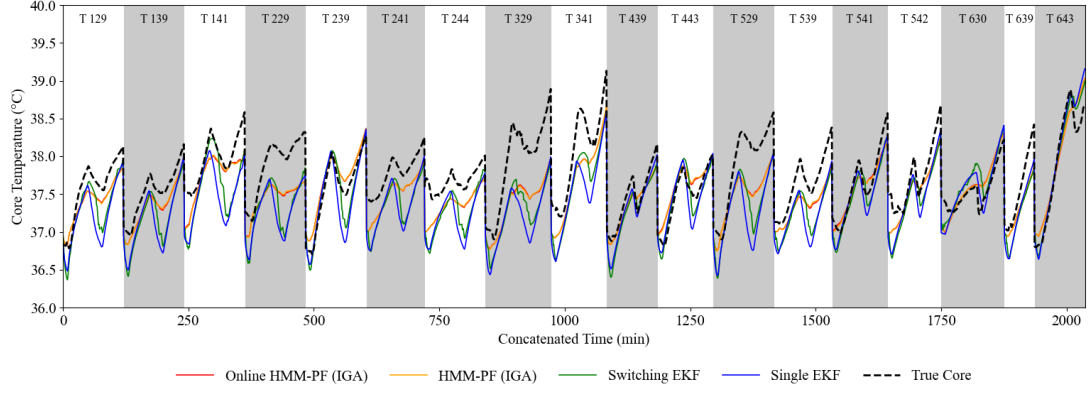
Method	RMSE	MAE	R^2	MD	SD
Online HMM-PF (IGA)	0.371	0.307	0.475	-0.109	0.355
HMM-PF (IGA)	0.372	0.309	0.473	-0.112	0.354
Online HMM-PF (Standard)	0.389	0.324	0.424	-0.088	0.379
HMM-PF (Standard)	0.389	0.324	0.423	-0.089	0.379
Switching PF (Standard)	0.451	0.382	0.224	-0.056	0.448
Single PF (Standard)	0.467	0.396	0.170	-0.121	0.451
Online HMM-EKF	0.468	0.394	0.165	-0.003	0.468
HMM-EKF	0.469	0.395	0.160	-0.004	0.469
Switching EKF	0.543	0.461	-0.126	+0.010	0.543
Single EKF	0.638	0.534	-0.549	+0.075	0.633

along the time axis. The black line denotes the reference core temperature, while colored lines represent predictions from selected models. For clarity, only four representative models are included: the two best-performing ones (Online HMM-PF (IGA) and HMM-PF (IGA)) and two EKF-based baselines (Switching EKF and Single EKF). Alternating grey and white background bands are used to visually separate individual trials. Each trial is labelled using a three-digit identifier, where the first digit indicates the experimental condition and the final two digits correspond to the original participant number.

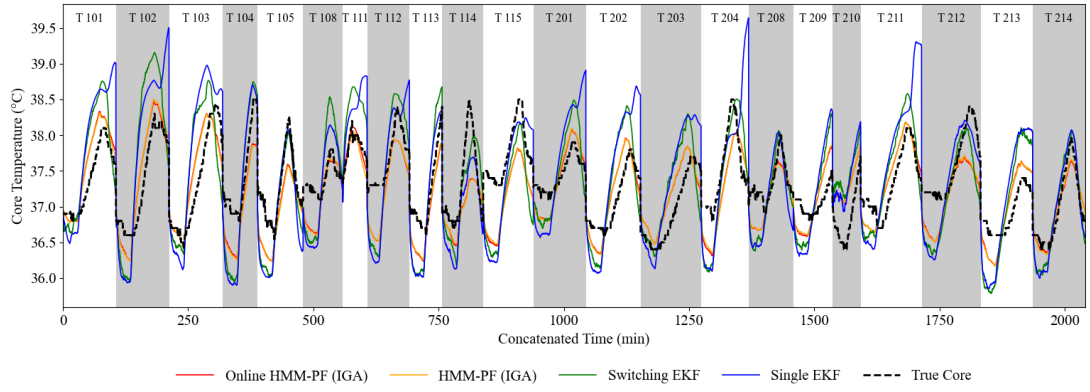
This visualization provides a holistic view of each model’s ability to track core temperature dynamics across varying sessions and physiological conditions.

To assess the agreement between predictions and reference temperatures across the full range, Bland–Altman plots for four representative models are presented—Online HMM-PF (IGA), HMM-PF (IGA), Switching EKF, and Single EKF—in Figure 5.4 and Figure 5.5, corresponding to Dataset 1 and Dataset 2, respectively. These plots visualize the difference between predicted and reference CBT values against their mean, highlighting systematic bias and limits of agreement.

By comparing specific model variants, the contribution of each component in the proposed framework can be isolated. A consistent performance advantage is observed



(a) Dataset 1



(b) Dataset 2

Figure 5.3: Predicted core temperature over time for (a) selected test trials from Dataset 1 and (b) all test trials from Dataset 2. The black line shows the reference core temperature. Colored lines correspond to predictions from Online HMM-PF (IGA), HMM-PF (IGA), Switching EKF, and Single EKF. Trials are concatenated and visually separated by alternating grey and white background regions. Each trial is labelled with a three-digit identifier, in which the first digit denotes the experimental condition setting and the final two digits correspond to the original participant number.

for particle filter (PF)-based methods over extended Kalman filter (EKF)-based ones. Across both datasets, even the simplest PF variant, such as the Single PF (Standard), achieves lower error and higher R^2 compared to most of the advanced EKF variants like the Online HMM-EKF. This difference is particularly pronounced on Dataset 2, where both the Switching EKF and Single EKF models yielded negative R^2 values, indicating predictions worse than the mean-value baseline.

Models that incorporate latent physiological states also show a clear advantage over those with a fixed observation model. For example, on Dataset 1, the HMM-PF (Standard) model achieves a lower RMSE (0.367°C) compared to the Single PF (Standard) model (0.383°C). This highlights the importance of capturing transitions

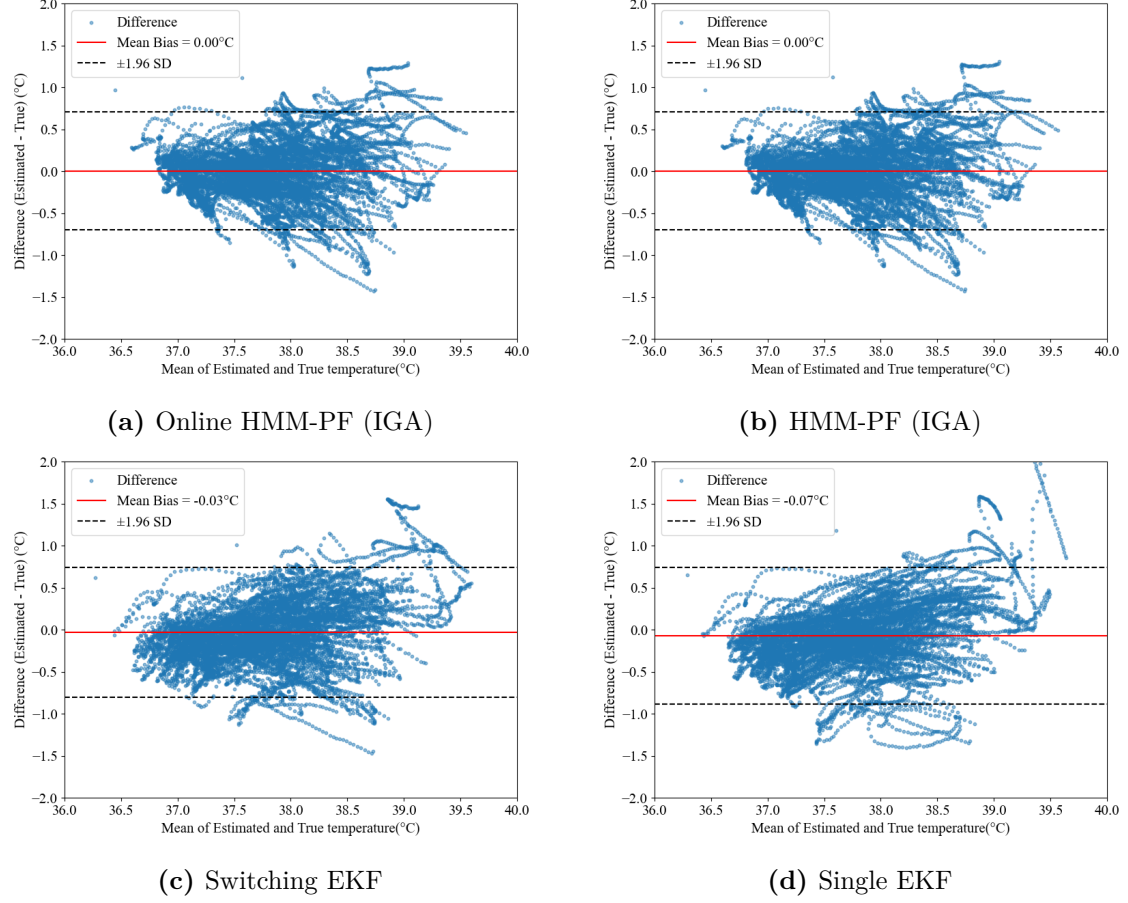


Figure 5.4: Bland–Altman plots of core temperature estimation on Dataset 1 for four representative models. Each subplot displays the prediction error (estimated – true) against the mean temperature. The solid line represents the average bias, and the dashed lines indicate the 95% limits of agreement.

between distinct physiological states, such as exercise and recovery, for accurate core temperature estimation.

Further improvements are achieved by incorporating the improved genetic algorithm (IGA) resampling strategy. On Dataset 1, the RMSE of HMM-PF decreases from 0.367°C under standard resampling to 0.359°C with IGA and tuned hyper-parameters. This result supports the value of adaptive resampling and automated parameter tuning in reducing estimation error.

Finally, introducing online adaptation of the HMM transition matrix yields subtle but consistent performance gains. On both datasets, the Online HMM variants showed better performance than their fixed-HMM counterparts in terms of both RMSE and R^2 . For instance, Online HMM-PF (Standard) performs better than HMM-PF (Standard) on Dataset 2. This suggests that real-time adaptation of

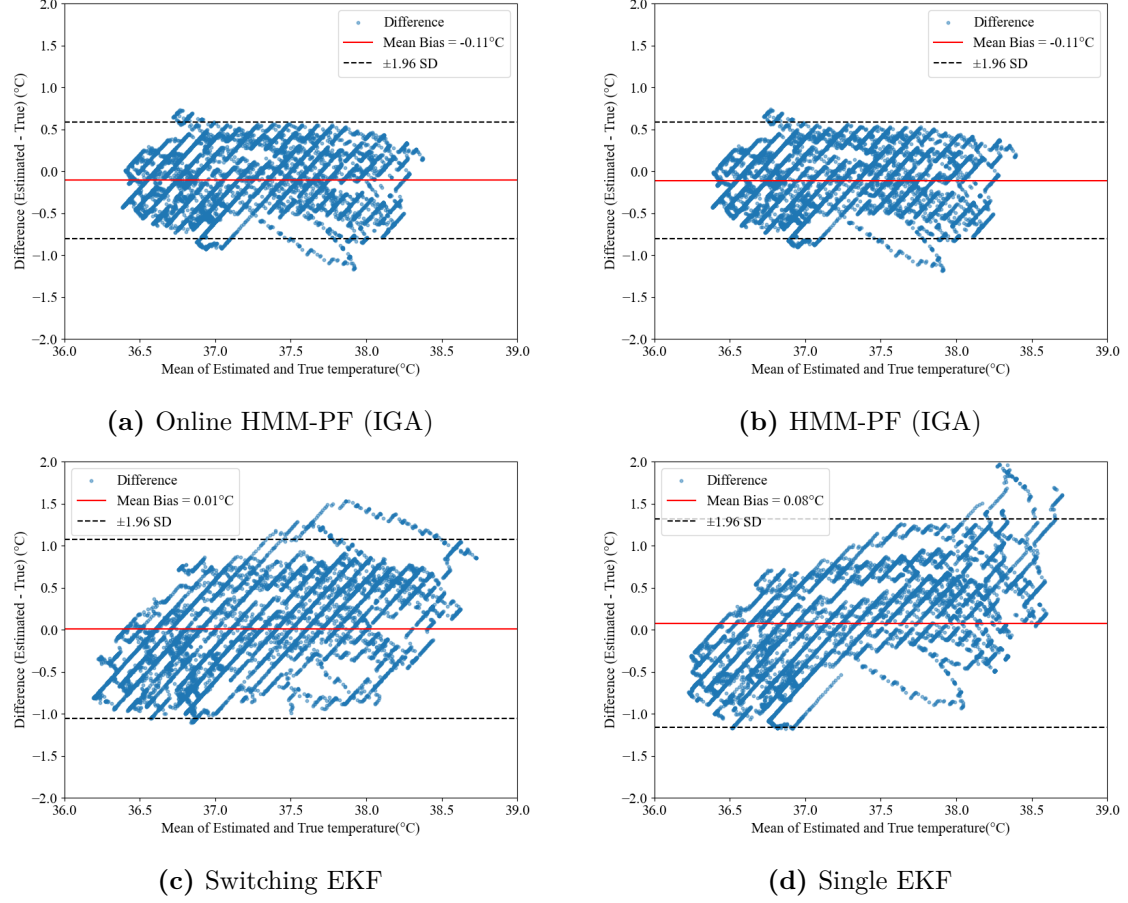


Figure 5.5: Bland–Altman plots of core temperature estimation on Dataset 2 for four representative models. Each subplot displays the prediction error (estimated – true) against the mean temperature. The solid line represents the average bias, and the dashed lines indicate the 95% limits of agreement.

transition dynamics helps improve robustness to individual variability and changing physiological conditions.

5.4 Discussion

In this study, a sophisticated Bayesian filtering framework was developed specifically to address the challenges of real-world physiological monitoring using wearable sensors. The goal was to enable accurate, non-invasive, real-time estimation of CBT using only the HR data available from common wearable devices. The findings clearly demonstrate that a Particle Filter (PF) framework, enhanced with a probabilistic state-fusion mechanism via an online Hidden Markov Model (HMM) and an advanced Improved Genetic Algorithm (IGA) for resampling, yields significantly more accurate and robust results than simpler baseline models. The proposed Online HMM-PF

(IGA) model consistently achieved better performance than all other variants across two distinct physiological datasets collected during exercise. This section interprets these findings, contextualizes them within the existing literature, acknowledges the study’s limitations, and suggests directions for future research.

The better performance of the proposed method stems from a multi-faceted approach that addresses several core challenges in physiological state estimation. A primary finding is the substantial performance gap between Particle Filter (PF)-based and Extended Kalman Filter (EKF)-based methods. This discrepancy does not arise from differences in the assumed observation noise—both filters adopt Gaussian observation models—but from how they approximate the posterior distribution. The EKF method approximates this distribution as a unimodal Gaussian and updates it through local linearization, which can fail in the presence of strong nonlinearity or non-Gaussian dynamics [220, 221]. In contrast, the particle filter maintains a flexible representation by tracking a whole population of possible temperature states at once. This approach is better suited for complex physiological systems, especially during ambiguous transitional phases like recovery. The PF can represent this uncertain relationship between heart rate and core temperature by exploring multiple plausible temperature trajectories simultaneously, a flexibility that the single-guess approach of the EKF lacks.

The results strongly support the importance of incorporating latent physiological states into the model. As illustrated in Figure 5.2, the relationship between HR and CBT differs systematically between the exercise and recovery phases. Data points during exercise tend to cluster tightly along a rising trend, indicating a more consistent HR–CBT relationship driven by increasing metabolic heat production [227]. In contrast, recovery-phase data exhibit greater variability and a flatter trend, likely reflecting complex and variable cooling processes such as vasodilation and heat dissipation [228]. Models that do not differentiate between these states—such as the Single EKF and Single PF—must fit a single function across these divergent regimes, resulting in compromised accuracy and degraded performance.

Furthermore, the method used for state-switching plays a critical role in overall model performance. While simple rule-based “hard-switching” improves upon using

a single model, it remains brittle during periods of signal ambiguity or transition. In contrast, probabilistic “soft-switching” with a Hidden Markov Model (HMM) offers greater robustness by maintaining a belief distribution over possible physiological states. This probabilistic fusion enables smoother transitions and more stable estimations across both filtering frameworks. As shown in Tables 5.2 and 5.3, the HMM-based PF models consistently achieved lower RMSE and higher R^2 than the Switching PF model. A similar trend was observed within the EKF framework, where the HMM-EKF models also achieved lower RMSE and higher R^2 than the Switching EKF model on both datasets.

Notably, this advantage is not limited to the Particle Filter. Within the EKF family, the HMM-EKF (RMSE 0.380 °C) also showed lower RMSE than the Switching EKF (RMSE 0.395 °C) on Dataset 1, underscoring that the benefits of soft state modeling are fundamental to the estimation task itself rather than specific to the choice of filtering algorithm.

It is important to recognize the trade-offs involved. Although PF-based methods offer better accuracy, they incur higher computational costs. In contrast, EKF-based models—particularly those incorporating HMM-based switching—provide a reasonable compromise between accuracy and efficiency. For resource-constrained devices such as low-power wearables, combining EKF with an online-updated HMM may offer a practical and lightweight solution, retaining much of the state modeling benefit without the full overhead of particle-based inference.

The use of an advanced resampling strategy proved valuable within the PF framework [229]. The performance gain of the IGA-equipped models over those with standard resampling highlights the practical challenge of particle impoverishment. By actively maintaining particle diversity through its genetic operators, the IGA leads to more stable and accurate state tracking, as evidenced by the consistent reduction in RMSE when moving from a standard PF to an IGA-based PF (e.g., on Dataset 1, HMM-PF (Standard) vs. HMM-PF (IGA)).

In addition, the consistent, albeit modest, improvement from online HMM adaptation underscores its potential value. This subtlety is likely due to the structured nature of the laboratory protocols where activity patterns are relatively uniform.

However, the demonstrated ability to fine-tune transition probabilities in real time is a crucial feature for future real-world applications where user behavior is unpredictable and highly personalized, making the online framework inherently more robust.

The study provides a direct performance comparison to prior work by reimplementing two foundational EKF-based approaches on two exercise datasets designed to capture physiological responses under heat stress: the method of Buller et al. [46, 169], which corresponds to the Single EKF, and the method of Rizvi et al. [138], which corresponds to the Switching EKF. The results in Table 5.2 and Table 5.3 quantify the benefits of the proposed enhancements over these established baselines. On Dataset 1, the Single EKF yielded an RMSE of 0.421 °C. The Switching EKF improved upon this, achieving an RMSE of 0.395 °C. The proposed Online HMM-PF (IGA) model, however, further reduced this error to 0.359 °C. The performance gap was even more pronounced on the more challenging Dataset 2, where both the Single EKF and Switching EKF failed ($R^2 < 0$), while the model maintained robust accuracy (RMSE = 0.371 °C). By replacing the deterministic switching rule with a probabilistic HMM, employing a non-parametric filter (PF), and enhancing it with an adaptive resampling algorithm (IGA), the framework demonstrates a clear and quantifiable advancement in estimation accuracy.

Despite the promising results, this study has several limitations that offer valuable directions for future research. First, the use of a two-state Hidden Markov Model (HMM) to represent latent physiological states (exercise and recovery) is a simplification. A more granular, multi-state formulation may better capture the continuum of physiological transitions, such as warm-up or passive cooling phases.

Second, the state-specific observation models (h_0 and h_1) were trained on data partitioned using the same rule-based heuristic as employed by the baseline Switching models. While this strategy is common and practical, it introduces a potential source of bias—particularly in transition periods—where heuristic mislabeling can propagate into the learned models. This limitation may inherently cap the performance ceiling of the HMM-based fusion methods. Future work could investigate semi-supervised or unsupervised approaches to jointly infer latent states and train observation

models from unlabeled data, reducing reliance on rigid heuristics and improving robustness in ambiguous regimes.

Third, the proposed model in this study only used the HR as the input and did not incorporate exogenous signals. Augmenting the model with additional inputs—such as accelerometer-derived workload or ambient thermal load—could improve its ability to capture dynamic physiological responses, especially under free-living conditions.

Furthermore, it is recognized that for predicting severe outcomes like exertional heat stroke, the core-to-skin temperature gradient is a physiologically critical variable [230]. The study focuses exclusively on providing a high-fidelity estimate of CBT, which we propose as a foundational prerequisite for any model that would seek to calculate this gradient non-invasively. Future work should therefore aim to integrate the robust CBT estimator with wearable skin temperature sensors to provide a more comprehensive view of an individual’s thermoregulatory status.

Finally, while the framework was rigorously evaluated on two well-controlled laboratory datasets, generalization to real-world settings remains an open question. Future studies should test the system in uncontrolled, real-world environments using consumer-grade wearables, where motion artifacts, variable adherence, and environmental unpredictability pose new challenges. Such validation is essential to ensure the model’s practical utility in continuous health monitoring and early detection of heat-related stress.

The importance of the findings lies in addressing a critical implementation gap in exertional heat illness prevention and management. While rectal thermometry remains the definitive method for diagnosing life-threatening exertional heat stroke, its invasive nature makes it unsuitable for the continuous, field-based monitoring required for proactive intervention. Existing environmental monitoring strategies are also inadequate, as they overlook the individual physiological variability that often precipitates exertional heat illness, helping explain its persistently high incidence [12, 190, 231].

At the same time, commercial wearable sensors are already widely used to capture external performance metrics, yet estimating internal physiological states such as CBT has remained a major challenge. Foundational HR-based models have shown feasibility but often lose robustness during the ambiguous and dynamic

periods common in athletic activity. By introducing a probabilistic, non-parametric framework that explicitly accounts for signal uncertainty, the work overcomes these limitations and achieves stable performance even during transitions such as the shift from intense exercise to recovery. This demonstrated enhancement in robustness and accuracy underscores that the intelligence lies not only in the sensors themselves but in physiologically aware algorithms that process their data. In this way, ubiquitous HR sensing can be transformed into a scalable biosensing technology for sports and occupational safety, enabling practical real-time monitoring to support performance, training safety, and injury prevention.

5.5 Conclusion

In this study, a novel framework for analyzing wearable heart rate sensor data was introduced, demonstrating that core body temperature can be accurately estimated in real time using only heart rate. By integrating probabilistic state-switching to capture exercise and recovery dynamics with a non-parametric particle filter to handle signal uncertainty, the results show that a physiologically motivated approach is essential for robust performance. The proposed model achieved lower estimation errors than the evaluated EKF-based methods across two exercise datasets designed to capture physiological responses under heat stress. The framework was specifically developed for heat-stress monitoring in hot environments, where continuous and minimally intrusive estimation of core body temperature is particularly important for applications such as sports and occupational safety. Future studies could further investigate whether the proposed approach can be extended to cold environments. This work highlights a scalable pathway to augment common wearable devices with a critical new sensing capability, and future validation in free-living conditions will be key to confirming its utility for sports, occupational safety, and broader health monitoring.

6

Estimating Core Body Temperature from Heart Rate Using a Residual-Compensated Adaptive Kalman Filter

Chapters 4 and 5 addressed the non-linear relationship between heart rate and CBT by segmenting physiological dynamics into discrete states (e.g., exercise and recovery), using variance-based fusion and probabilistic inference, respectively. However, systematic errors can persist even within correctly identified states due to sensor noise, model simplifications, and individual physiological variability.

To address these challenges, this chapter investigates an alternative micro-level residual compensation strategy. Instead of switching or fusing between multiple models, this study proposes a Residual-Compensated Adaptive Kalman Filter (RCAKF). This framework integrates a lightweight Long Short-Term Memory (LSTM) network directly into the observation equation of a single EKF to compensate the observation model in real-time. Additionally, an adaptive noise estimator dynamically adjusts for measurement uncertainty.

Evaluated on a controlled dataset of 22 participants (the same Dataset 1 used in Chapter 5), the RCAKF achieved higher accuracy (RMSE = 0.31 °C) compared to baseline models, including standard EKF (RMSE = 0.39 °C), BKFB (RMSE = 0.48 °C), MAKF (RMSE = 0.38 °C) and standalone LSTM networks (RMSE = 0.45 °C).

These results demonstrate that hybrid error correction leads to a strong improvement in estimation accuracy compared with unaugmented Kalman filtering. The integration of deep learning with Kalman filtering supports the hybrid modeling

concept and forms the methodological basis for the physics-guided hybrid framework developed in Chapter 7.

6.1 Introduction

Core body temperature is a critical physiological parameter reflecting the thermal homeostasis of the human body. It plays a particularly important role in preventing heat stress-related injuries during strenuous physical activity, occupational labor, or exposure to high ambient temperatures [186, 232, 233]. Under such conditions, increased metabolic heat production and reduced heat dissipation can cause core temperature to rise rapidly. If core temperature exceeds 41 °C within a short period, it can result in severe thermal injury, including protein denaturation and cellular necrosis. Without timely intervention, further increases in body temperature may lead to heat stroke, a medical emergency with a mortality rate of approximately 12% among adults [185]. According to reports, an average of 5,946 individuals are treated for exertional heat-related illness (EHRI) in emergency departments across the United States each year [234]. Therefore, accurate monitoring and early warning of core body temperature during exercise, training, or high-risk work environments is essential for preventing injury and ensuring physiological safety.

Although rectal temperature is often considered too invasive for routine or continuous use, it remains the criterion standard for assessing core temperature in hyperthermic individuals [20, 30]. As recommended by the National Athletic Trainers' Association (NATA) position statement, rectal temperature offers valid and reliable measurements both at rest and during exercise, and is considered an appropriate reference method for evaluating core body temperature in physically active settings [15]. In practice, however, peripheral measurement sites—such as the axilla, tympanic membrane, oral cavity, and forehead—are often used for convenience [2, 32]. These sites, while more accessible, have been shown to produce substantial mean bias compared to rectal temperature, particularly during exertional heat stress. Therefore, they are generally inadequate for real-time or high-precision core temperature monitoring [36, 37].

In parallel, estimating core temperature directly from readily accessible physiological signals—such as heart rate and skin temperature—has emerged as a more scalable strategy for unobtrusive, real-time monitoring in athletic and occupational settings [50, 138, 146, 147, 169]. This approach is particularly valuable in scenarios requiring continuous and unobtrusive monitoring, including athletic performance tracking, heat acclimatization, and occupational safety.

A variety of modeling approaches have been developed based on wearable sensor data, which can be broadly categorized. Firstly, purely data-driven models typically rely on supervised learning algorithms to capture nonlinear mappings between multimodal sensor inputs and core temperature [152]. For example, Niedermann et al. achieved an RMSE of approximately 0.28 °C to 0.32 °C using multiple linear regression with up to six signals, including heart rate, skin temperature and heat flux [49]. While often accurate, these “black-box” models can require extensive sensor arrays and may lack physiological interpretability.

Secondly, hybrid physics-based models offer a more interpretable alternative by integrating real-time sensor data into a physiological framework. For instance, a recent compartment model proposed by Hirata et al., driven by heart rate and ambient conditions, achieved a high accuracy with an RMSE as low as 0.14 °C by using the heart rate data to estimate metabolic rate within a simplified thermodynamic structure [235]. While offering physiological interpretability, these models can still require multiple inputs and carefully tuned parameters.

Thirdly, in contrast to multi-sensor approaches, several studies have explored sequential estimation frameworks that require only minimal physiological inputs. State-space models—particularly Kalman filtering—offer a principled approach by treating core body temperature as a latent state that evolves over time [46, 170]. Buller et al. were the first to propose an extended Kalman filter (EKF) for core temperature estimation using heart rate as the sole observation input [46], and their model has since been widely adopted and validated across a range of physical activities and environmental conditions [137, 170, 173]. By relying exclusively on heart rate, these methods maximize sensor simplicity, enabling practical deployment in wearable devices and field settings where minimizing hardware complexity, energy consumption,

and cost is critical. This sensor-efficient design also facilitates applications in large-scale monitoring, athletic training, occupational health, and circadian rhythm analysis. Building upon this foundation, Buller’s group further refined their model to improve its accuracy during periods of rest by replacing the original quadratic observation model with a generalized logistic (sigmoid) function [219]. This modification better captured the saturation behaviour of heart rate at both low and high core temperatures, thereby reducing systematic underestimation at rest and making the model more suitable for continuous monitoring of circadian variations in core body temperature.

Recent developments in EKF-based core temperature estimation can be broadly grouped under a paradigm we refer to as Macro-Strategy Switching. This approach enhances model adaptability by employing multiple pre-defined EKF models, each designed for a specific physiological phase. For instance, the method proposed by Rizvi and Udayraj switches models based on a moving average of heart rate trends [138], while the more recent Biphasic Kalman Filter-Based (BKFB) model by Falcone et al. uses a simple heart rate threshold to switch between ‘exercise’ and ‘recovery’ filters [148]. While these approaches offer interpretable, computationally efficient solutions that improve upon a single static model, their switching is discrete and reactive. They remain fundamentally incapable of addressing the core challenge of non-stationary conditions: the dynamic, time-varying errors inherent in the observation model itself, which arise from relying on fixed-form equations.

To mitigate these limitations, in this study, a complementary paradigm is explored: Micro-Level Real-Time Correction. Hybrid models that integrate Kalman filters with deep learning architectures like LSTM have demonstrated considerable success in various engineering and environmental science domains. For example, they have been employed to enhance the accuracy of streamflow forecasting in hydrology [236], to improve state estimation in vehicle dynamics control [237], and to track non-linear trajectories of underwater targets [238].

Inspired by these successes, this work pioneers the application of such a hybrid framework to the distinct challenge of core body temperature estimation. In this work,

the Residual-Compensated Adaptive Kalman Filter (RCAKF) is introduced as a concrete implementation of this paradigm. The core philosophy of this approach is not to alternate between models, but to augment a single EKF with mechanisms that enable continuous adaptation. Specifically, the RCAKF integrates a long short-term memory (LSTM) network to model and compensate for structured, time-varying residuals in the observation process, alongside an adaptive noise estimator that dynamically adjusts measurement uncertainty. This hybrid design retains the interpretability and structure of classical Kalman filtering while leveraging the representational power of deep learning. Operating solely on heart rate input, RCAKF offers enhanced flexibility and improved accuracy under diverse physiological conditions.

This study makes three primary contributions. First, the micro-level real-time correction paradigm is formalized and validated in the context of non-invasive core body temperature estimation using heart rate, and its advantages over existing macro-level strategy-switching methods are demonstrated. Second, the Residual-Compensated Adaptive Kalman Filter (RCAKF) is developed as a hybrid filtering framework that integrates LSTM-based residual compensation into an adaptive EKF architecture. Third, comprehensive validation demonstrates that RCAKF achieved better performance than established baselines, highlighting its practical value for real-world wearable applications.

6.2 Methods

6.2.1 Dataset

This study utilized a controlled laboratory dataset comprising 22 healthy adult participants who performed treadmill-based exercise under a range of thermal conditions [150, 151]. The same dataset was used as Dataset 1 in Chapter 5. A detailed description of the participants, experimental protocol, and measurements is provided in Table 5.1. Heart rate values were obtained directly from the chest ECG measurement system.

6.2.2 Extended Kalman Filter (EKF)

To enable non-invasive estimation of core body temperature from heart rate, this study employs an extended Kalman filter (EKF) framework in which core temperature is treated as a latent state and heart rate as the observed signal [137, 173]. The EKF recursively updates the temperature estimate by combining a dynamic time update model and a nonlinear observation model, both empirically derived from training data.

The *time update model* describes the evolution of core temperature over time and is implemented as a first-order autoregressive process learned via linear regression. This model captures the gradual physiological change in temperature and introduces process noise to reflect uncertainty.

The *observation model* establishes the mapping from core temperature to heart rate using a second-order polynomial regression, reflecting their nonlinear physiological relationship. Observation noise is modeled to account for deviations between predicted and actual heart rate.

Together, these models allow the EKF to balance prior knowledge with real-time heart rate measurements, enabling sequential, real-time estimation of core body temperature from wearable signals.

State Transition Model

$$T_t = a_1 T_{t-1} + a_0 + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \gamma^2) \quad (6.2.2.1)$$

Here, T_t is the latent core temperature at time t , and T_{t-1} is the previous estimate. The parameters a_1 and a_0 are obtained from linear regression, and γ^2 represents the process noise variance.

Observation Model

$$\hat{H}R_t = b_2 T_t^2 + b_1 T_t + b_0 + \eta_t, \quad \eta_t \sim \mathcal{N}(0, \sigma^2) \quad (6.2.2.2)$$

In this model, $\hat{H}R_t$ is the predicted heart rate corresponding to core temperature T_t . The coefficients b_2 , b_1 , and b_0 are obtained from second-order polynomial regression. The term η_t is Gaussian observation noise with variance σ^2 .

The recursive filtering process proceeds as follows:

Prediction Step

$$\hat{T}_t = a_1 T_{t-1} + a_0 \quad (6.2.2.3)$$

$$\hat{v}_t = a_1^2 v_{t-1} + \gamma^2 \quad (6.2.2.4)$$

The prior temperature estimate $\hat{T}_t$ and its associated variance $\hat{v}_t$ are predicted using the state transition model.

Observation Prediction

$$\hat{H}R_t = b_2 \hat{T}_t^2 + b_1 \hat{T}_t + b_0 \quad (6.2.2.5)$$

The expected heart rate is computed from the predicted core temperature.

Extended Kalman filter mapping function variance coefficient

$$m_t = \frac{d\hat{H}R_t}{dT} = 2b_2 \hat{T}_t + b_1 \quad (6.2.2.6)$$

To enable Kalman-based updates with a nonlinear observation model, the function is linearized and the term m_t denotes the extended Kalman filter mapping function variance coefficient.

Kalman Gain

$$k_t = \frac{\hat{v}_t \cdot m_t}{m_t^2 \cdot \hat{v}_t + \sigma^2} \quad (6.2.2.7)$$

The Kalman gain k_t determines the weight given to the new observation in updating the state estimate.

Update Step

$$T_t = \hat{T}_t + k_t \cdot (HR_t - \hat{H}R_t) \quad (6.2.2.8)$$

The posterior temperature estimate is corrected based on the observed heart rate HR_t .

Variance Update

$$v_t = (1 - k_t m_t) \cdot \hat{v}_t \quad (6.2.2.9)$$

The variance is updated to reflect the new estimation uncertainty.

The EKF is initialized with a core temperature of $T_0 = 37.0^\circ\text{C}$ and a variance of $v_0 = 0.01\text{ }^\circ\text{C}^2$ in this study. All model parameters are fitted from training data and fixed during inference. This filter offers a lightweight and interpretable solution for real-time estimation of core body temperature using heart rate data alone.

6.2.3 Residual-Compensated Adaptive Kalman Filter (RCAKF)

To improve the robustness of core temperature estimation in the presence of time-varying and non-Gaussian noise, a Residual-Compensated Adaptive Kalman Filter (RCAKF) is proposed. This method extends the standard extended Kalman filter (EKF) by incorporating a deep learning-based residual compensation module and an adaptive observation noise estimator.

Specifically, the RCAKF augments the polynomial observation model with a data-driven correction term predicted by a long short-term memory (LSTM) network. Additionally, the observation noise variance is no longer assumed fixed but is dynamically estimated from recent prediction errors via a sliding window approach.

Observation Model with Residual Correction To account for systematic, state-dependent deviations in the predicted heart rate not captured by the polynomial model, a learned residual correction term $\hat{r}_t$ is introduced, which is estimated by an LSTM network:

$$\hat{r}_t = \text{LSTM}(HR_{t-n:t-1}) \quad (6.2.3.1)$$

The corrected observation becomes:

$$\hat{H}R_t = b_2 \hat{T}_t^2 + b_1 \hat{T}_t + b_0 + \hat{r}_t \quad (6.2.3.2)$$

This addition allows the observation model to dynamically adapt its output based on recent heart rate patterns, which is crucial for handling non-stationary physiological states.

Prediction Error and Adaptive Noise Estimation The prediction error at time t is computed as:

$$\epsilon_t = HR_t - \hat{H}R_t \quad (6.2.3.3)$$

Furthermore, to account for time-varying measurement uncertainty, the fixed observation noise assumption is replaced by adaptively estimating the variance $\hat{\sigma}_t^2$ from a sliding window of recent prediction errors, where a buffer $\mathcal{S} = \{\epsilon_{t-w+1}^2, \dots, \epsilon_t^2\}$ of length w is maintained and used for variance estimation:

$$\hat{\sigma}_t^2 = \max\left(0, \text{mean}(\mathcal{S}) - m_t^2 \hat{v}_t\right) \quad (6.2.3.4)$$

where $m_t = \frac{d\hat{H}R_t}{dT} = 2b_2\hat{T}_t + b_1$ is the linearized observation model derivative, and $\hat{v}_t$ is the predicted state variance. The buffer provides a short-term history of recent prediction errors, allowing the observation noise estimate to adapt to gradual changes in measurement quality while reducing sensitivity to transient fluctuations or isolated outliers.

Kalman Gain and State Update The Kalman gain is computed using the adapted noise variance:

$$k_t = \frac{\hat{v}_t \cdot m_t}{m_t^2 \cdot \hat{v}_t + \hat{\sigma}_t^2} \quad (6.2.3.5)$$

$$T_t = \hat{T}_t + k_t \cdot \epsilon_t \quad (6.2.3.6)$$

$$v_t = (1 - k_t m_t) \cdot \hat{v}_t \quad (6.2.3.7)$$

The main steps of the proposed RCAF method are summarized in Algorithm 2.

Implementation Details To ensure compatibility with wearable deployment constraints, the residual model is intentionally designed to be lightweight. It consists of a single-layer LSTM network with a hidden dimension of 32, trained using the Adam optimizer with a learning rate of 0.001. Early stopping is applied with a patience of 5 epochs based on validation loss.

The model operates solely on a univariate input sequence of heart rate measurements, eliminating the need for multimodal signals or demographic information.

Algorithm 2 Residual-Compensated Adaptive Kalman Filter (RCAKF)

1. **Input:** Initial core temperature T_0 , variance v_0 ;
coefficients a_0, a_1, b_0, b_1, b_2 ;
process noise standard deviation γ ; initial observation noise standard deviation σ_0 ;
residual model LSTM; LSTM window size n ; adaptive window size w

2. Initialize squared error buffer: $\mathcal{S} \leftarrow [\sigma_0^2] \times w$

3. For each time step $t = 1$ to T , do:

Step 1: Predict core temperature and its variance:

$$\hat{T}_t = a_1 T_{t-1} + a_0, \quad \hat{v}_t = a_1^2 v_{t-1} + \gamma^2$$

Step 2: Estimate residual using LSTM:

$$\hat{r}_t = \text{LSTM}(H R_{t-n:t-1})$$

Step 3: Observation prediction with residual compensation:

$$\hat{H}R_t = b_2 \hat{T}_t^2 + b_1 \hat{T}_t + b_0 + \hat{r}_t, \quad m_t = 2b_2 \hat{T}_t + b_1$$

Step 4: Compute prediction error and update observation noise estimate:

$$\epsilon_t = H R_t - \hat{H}R_t$$

$$\mathcal{S} \leftarrow \mathcal{S}[1:] \cup \{\epsilon_t^2\}, \quad \hat{\sigma}_t^2 = \max(0, \text{mean}(\mathcal{S}) - m_t^2 \hat{v}_t)$$

Step 5: Compute Kalman gain and update:

$$k_t = \frac{\hat{v}_t m_t}{m_t^2 \hat{v}_t + \hat{\sigma}_t^2}$$

$$T_t = \hat{T}_t + k_t \cdot \epsilon_t, \quad v_t = (1 - k_t m_t) \cdot \hat{v}_t$$

4. **Output:** Estimated temperature sequence $\{T_t\}_{t=1}^T$
-

Specifically, the LSTM takes as input a sliding window of the past $n = 10$ heart rate values. Despite the input simplicity, this architecture is sufficient to capture structured deviations in the observation model under varying physiological conditions.

To evaluate model generalization across sessions and experimental conditions, a nested 3-fold cross-validation strategy based on trial-level splitting is adopted. Each trial corresponds to an independent data collection session performed under specific thermal, activity, and clothing conditions. Trials are randomly divided

into three disjoint folds, ensuring complete separation between training and testing trials. In each outer fold, an 80/20 split of the training set is used to construct an inner validation set for early stopping and adaptive parameter tuning. All model parameters—including autoregressive coefficients, polynomial observation mappings, and LSTM weights—are fitted exclusively using the outer training trials.

The adaptive variance window size w is selected by grid search over $\{5, 10, 20, 30, 50\}$ using inner validation performance. This configuration ensures that the residual-compensated adaptive Kalman filter (RCAKF) remains computationally efficient and deployable in real time on wearable platforms, while maintaining robust performance across unseen trial conditions.

The algorithms were implemented and evaluated using Python 3.10, NumPy 2.0.2, SciPy 1.13.1, scikit-learn 1.5.2, and TensorFlow 2.20.0. All analyses were conducted on a consumer laptop equipped with an AMD Ryzen 7 processor without GPU acceleration.

6.2.4 Evaluation Metrics

To assess model performance and generalization, a nested 3-fold cross-validation procedure based on complete experimental trials is employed. Each outer fold contained a disjoint subset of entire trials, ensuring that no partial session was used simultaneously for training and testing. This validation strategy evaluates the model’s ability to generalize across different thermal environments, clothing conditions, and physical activity patterns.

The model accuracy is assessed using three standard metrics: root mean squared error (RMSE), mean absolute error (MAE), and coefficient of determination (R^2). These metrics were computed between the estimated and reference core temperature values across all time points in the test sets.

In addition to standard error metrics, the Bland–Altman analysis is employed to evaluate the agreement between estimated and reference core temperatures [209].

The analysis reports the mean difference (MD), representing the average bias, and the standard deviation (SD) of the differences. The 95% limits of agreement, calculated as $MD \pm 1.96 \cdot SD$, indicate the interval within which most residuals are expected to fall under the assumption of normality [106].

To further assess each model’s ability to identify high-risk core temperature episodes, the receiver operating characteristic (ROC) analysis was conducted [239]. The threshold of 38.0°C was adopted, consistent with occupational heat stress screening guidelines [181]. Binary classification performance was quantified using the area under the ROC curve (AUC), sensitivity (true positive rate), and specificity (true negative rate).

These threshold-based metrics offer practical insights into the models’ applicability in early warning systems and real-time decision-making, under both conservative and high-risk definitions of elevated core temperature.

6.2.5 Comparison Methods

To evaluate the effectiveness of the proposed residual-compensated adaptive Kalman filter (RCAKF), its performance was compared with several representative baselines encompassing both classical filtering and learning-based approaches:

- **Extended Kalman Filter (EKF):** This baseline implements a standard extended Kalman filter using a single global state transition and observation model, learned from all training data. The observation model is a second-order polynomial mapping from core temperature to heart rate, and the filter does not adapt to state-dependent or time-varying residual patterns.
- **Improved ECTempTM(ECTemp-S):** Based on the refined ECTempTM algorithm by Looney and Buller [219], this method replaces the quadratic observation model with a generalized logistic (sigmoid) function that better captures the saturation behaviour of heart rate across the full temperature range. The sigmoid model parameters are fit to the training data via non-linear regression, and its derivative is used in the Kalman update step. This approach enhances estimation accuracy during resting periods while maintaining fidelity during exercise.
- **Biphasic Kalman Filter-Based Model (BKFB):** Adapted from Falcone et al. [148], this method employs two separate EKF models for the *exercise* and *rest* phases. Phase switching is performed using a heart rate threshold-based

rule: if three consecutive heart rate values exceed a predefined threshold, the exercise-phase model is applied; otherwise, the rest-phase model is used. The optimal threshold is selected from the range 100–120 bpm via cross-validation.

- **Moving Average Kalman Filter (MAKF):** This method was proposed by Rizvi and Udayraj [138] to improve temperature estimation across different physiological phases. It uses two separate extended Kalman filters: one for the *exercise* phase and another for the *recovery* phase, each with its own transition and observation models.

The active model is determined by the trend of a five-minute moving average of heart rate. Specifically, if the current moving average is lower than the previous one, the recovery-phase model is selected; otherwise, the exercise-phase model is used. This switching rule allows the filter to adapt to changes in cardiovascular response during physical activity and rest.

- **LSTM:** A deep learning model composed of a single-layer long short-term memory (LSTM) network with a hidden dimension of 32. The network architecture was kept identical to the LSTM module used within the proposed RCAKF framework to ensure a fair comparison by controlling for differences in model capacity. The model receives a 10-minute window of past heart rate values as input and directly outputs the predicted core temperature at the current time step.

The network architecture and input format are identical to those used in the RCAKF model; the only difference lies in the prediction target. Whereas RCAKF uses the LSTM to estimate observation residuals, the LSTM model here directly regresses the core temperature. This provides a non-recursive, data-driven reference.

All methods are evaluated using a nested 3-fold cross-validation procedure with disjoint experimental trials. Within the training folds, an inner 3-fold cross-validation is performed to select model-specific hyperparameters, such as the heart rate threshold in BKFB and the adaptive window size in RCAKF. All baseline models are retrained from scratch on the same dataset following the same pipeline, thereby ensuring a fair and direct comparison.

Model parameters are trained exclusively on the outer training folds, and all performance metrics are reported based on the outer test folds to ensure an unbiased and consistent comparison.

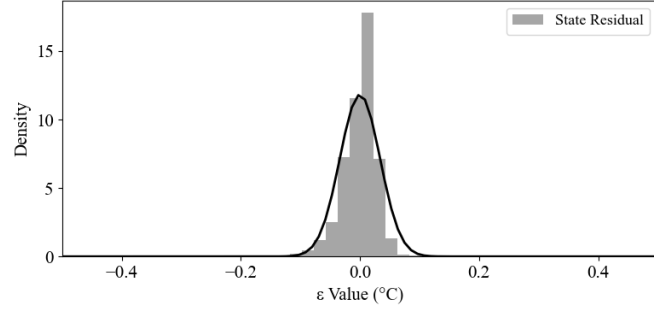
6.3 Results

Figure 6.1 illustrates the distribution of residuals from the state transition model and the observation model across all participants in the test set. Figure 6.1a shows the histogram of the prediction error $\epsilon_t = T_t - \hat{T}_t$ from the state transition model, which approximates a zero-mean Gaussian distribution. Figure 6.1b presents the distribution of observation residuals $r_t = HR_t - \hat{H}R_t$, comparing the fixed-form polynomial observation model with the LSTM-compensated version used in RCAKF. The residuals from the original observation model exhibit substantial skewness, deviating from Gaussianity. In contrast, the residuals after LSTM compensation are approximately zero-mean and symmetric, indicating that the residual correction mechanism effectively accounts for systematic, time-varying deviations in the observation process.

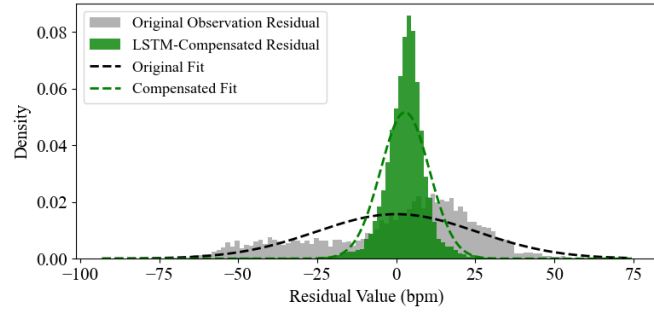
Figure 6.2 illustrates the relationship between core temperature and heart rate under different physiological states. The Before LSTM Compensation subplot shows the original mapping between core temperature and observed heart rate, revealing clear state-dependent deviations: heart rate tends to be underestimated during *Exercise* and overestimated during *Rest* and *Recovery*. These structured residuals reflect the limitations of using a fixed-form observation model across varying physiological conditions.

The After Compensation subplot presents the relationship after applying LSTM-based residual correction, where the compensated heart rate is computed by adding the predicted residual to the original model output. The corrected mapping exhibits improved alignment across states, with reduced bias and variance. This demonstrates that the RCAKF framework effectively accounts for state-dependent deviations by incorporating a data-driven compensation mechanism.

Taken together, Figures 6.1 and 6.2 reveal critical limitations of the fixed-form observation model in accurately capturing the dynamics of heart rate responses under



(a) State transition residuals (ϵ_t)



(b) Observation residuals (r_t)

Figure 6.1: Histogram of model residuals across the test set. (a) Residuals from the state transition model. (b) Residuals from the observation model.

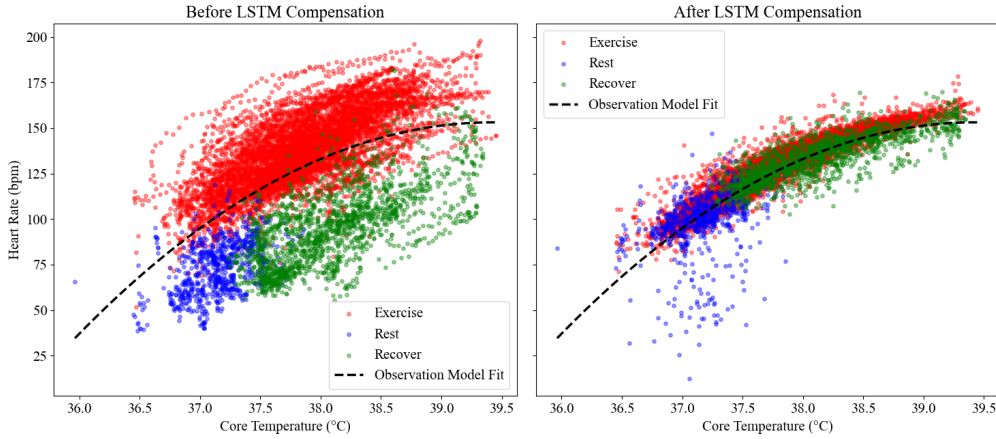


Figure 6.2: State-dependent relationship between core temperature and heart rate before and after LSTM-based residual compensation. Left: Original mapping between core temperature and observed heart rate, showing state-specific deviations. Right: Compensated mapping after adding LSTM-predicted residuals to the model output. Each point is color-coded by physiological state Rest, Recovery, or Exercise.

varying physiological states. While the state transition residuals largely conform to Gaussian assumptions, the observation residuals show systematic, state-dependent deviations—particularly during *Rest* and *Recovery* phases. These patterns underscore the need for a more flexible observation model capable of learning complex, temporal dependencies. The RCAF framework addresses this challenge by augmenting the standard EKF with a learned residual correction module, enabling improved core temperature estimation performance across diverse physiological conditions.

Figure 6.3 presents time-series comparisons between the estimated and reference core temperatures for selected test trials. Each colored line represents the prediction of a specific method, including EKF, ECTemp-S, BKFB, MAKF, LSTM, and the proposed RCAF. The black line denotes the reference core temperature. Trials are concatenated along the time axis, and alternating grey and white background regions are used to visually separate individual sessions. This figure provides a holistic view of each model’s ability to track temperature dynamics across diverse experimental conditions.

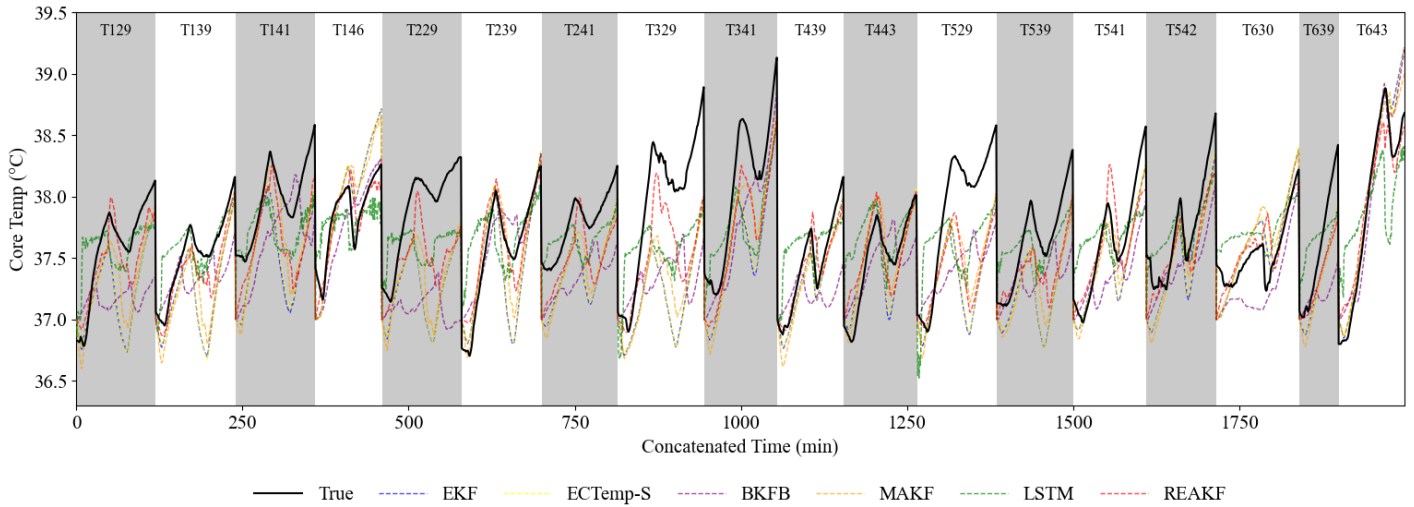


Figure 6.3: Time-series comparisons between the estimated and reference core temperatures for selected test trials. The black curve represents the reference core temperature, while colored curves correspond to different models, including EKF, ECTemp-S, BKFB, MAKF, LSTM, and the proposed RCAF. Alternating grey and white background regions separate individual trials for visual clarity. Each trial is labelled using a three-digit identifier, where the first digit denotes the experimental condition and the final two digits correspond to the original participant number

Figure 6.4 presents Bland–Altman plots comparing the predicted and reference

core temperatures. Each plot shows the mean difference (MD) and 95% limits of agreement ($MD \pm 1.96 SD$), highlighting the magnitude and consistency of model biases. The RCAF method demonstrates the narrowest spread and the smallest mean bias, indicating better agreement with the ground truth across the full temperature range.

Figure 6.5 shows normalized histograms of estimation errors ($\hat{T}_t - T_t$). The RCAF distribution is narrower and more centered, indicating improved accuracy and reduced bias.

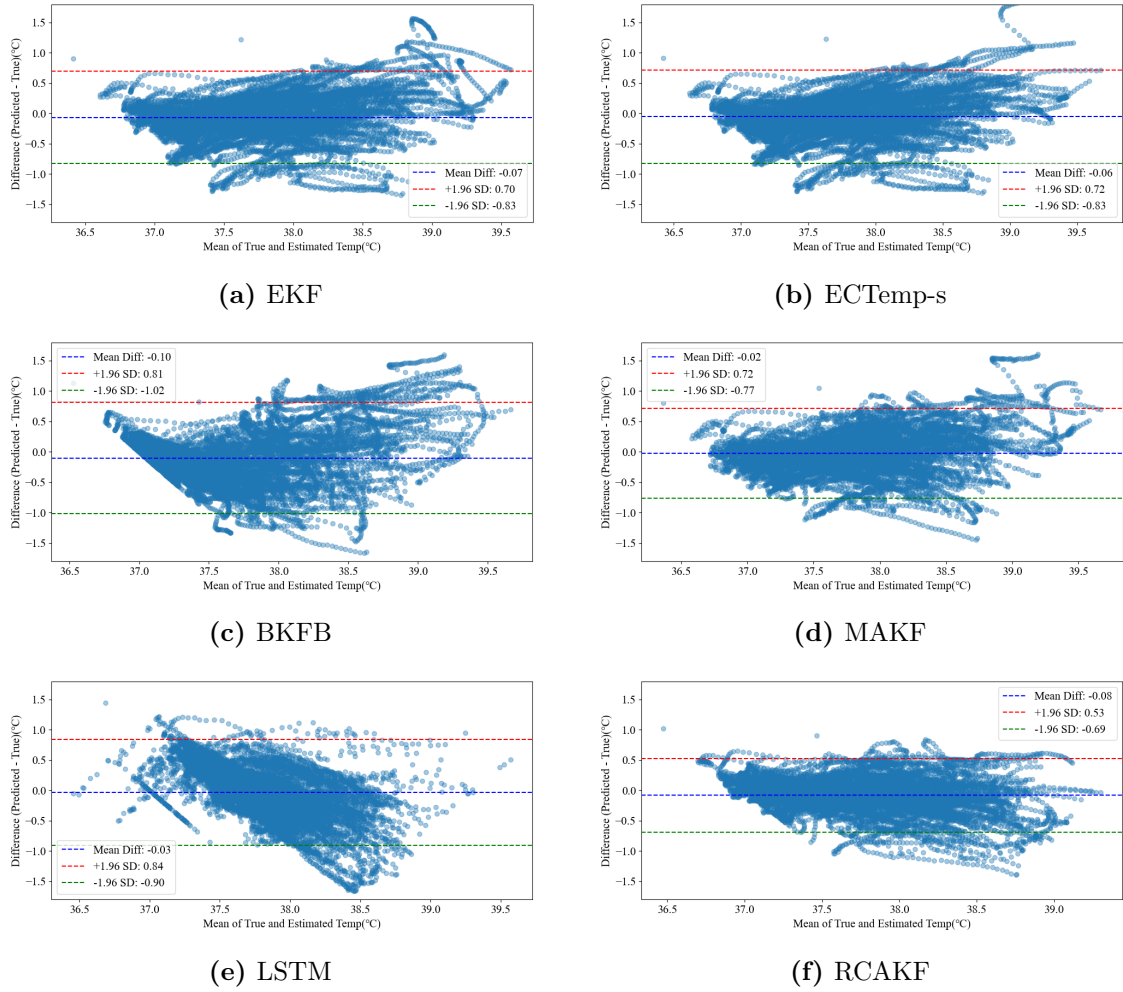
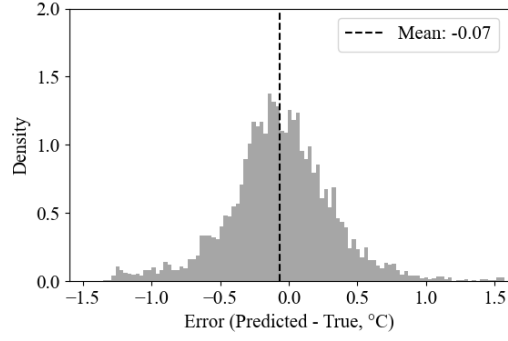
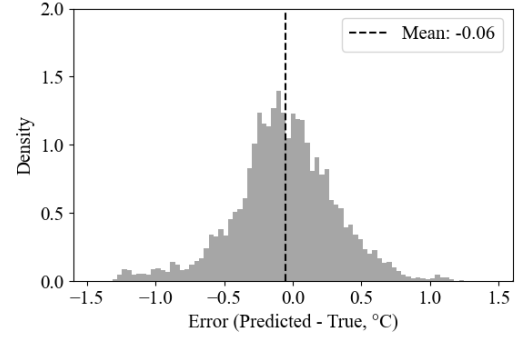


Figure 6.4: Bland–Altman plots comparing estimated and reference core temperatures for the six models: (a) EKF, (b) ECTemp-S, (c) BKFB, (d) MAKF, (e) LSTM, and (f) RCAF. Each plot shows the mean difference (blue line) and the 95% limits of agreement (green and red dashed lines) between predicted and ground-truth temperatures.

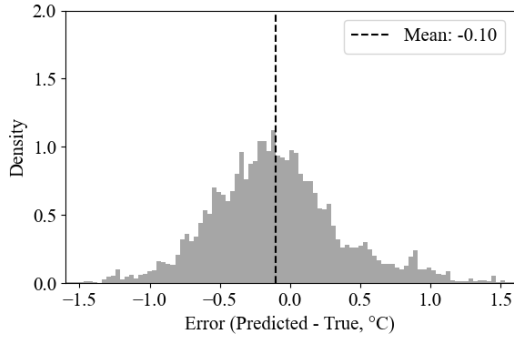
Figure 6.6 presents the receiver operating characteristic (ROC) curves for all six models under the 38.0°C threshold. The ROC curves illustrate each model’s ability to



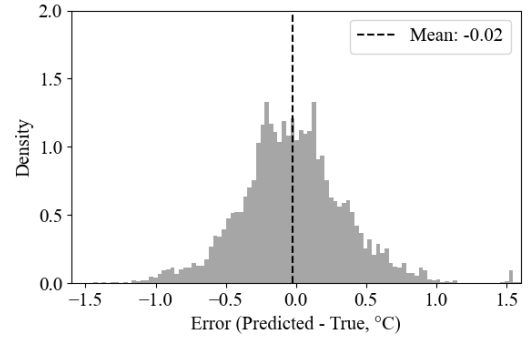
(a) EKF



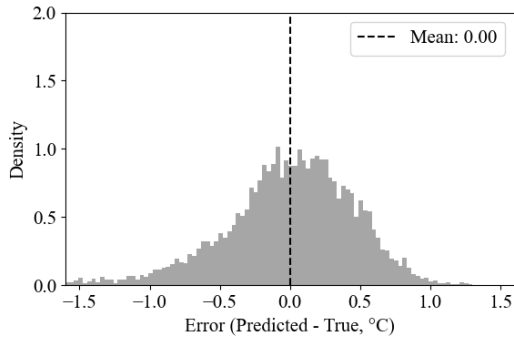
(b) ECTemp-S



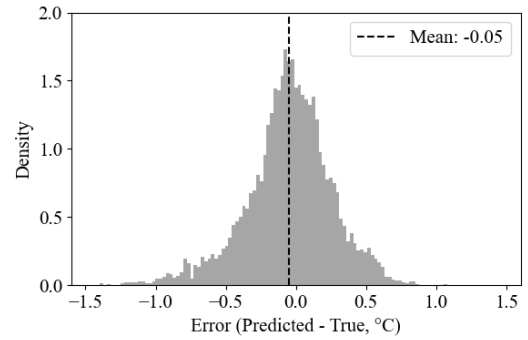
(c) BKFB



(d) MAKF



(e) LSTM



(f) RCAF

Figure 6.5: Normalized histograms of core temperature estimation errors ($\hat{T}_t - T_t$) for six models: (a) EKF, (b) ECTemp-S, (c) BKFB, (d) MAKF, (e) LSTM, and (f) RCAF. All histograms are plotted using the same bin size and range for visual comparability.

distinguish between episodes of normal and elevated core temperature ($\geq 38.0^{\circ}\text{C}$), with the area under the curve (AUC) quantifying their binary classification performance.

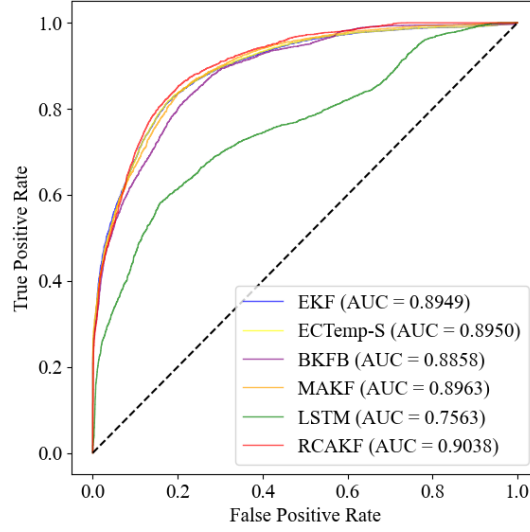


Figure 6.6: ROC curves for all six models (EKF, ECTemp-S, BKFB, MAKF, LSTM, and RCAF) in detecting elevated core temperature episodes ($\geq 38.0^{\circ}\text{C}$). The area under the curve (AUC) indicates the discriminative performance of each method.

Table 6.1 summarizes the comprehensive evaluation results for all six methods across the full dataset. The upper portion reports standard regression metrics, including root mean squared error (RMSE), mean absolute error (MAE), coefficient of determination (R^2), and mean difference $\pm$ standard deviation ($\text{MD} \pm \text{SD}$) from Bland–Altman analysis. The lower portion presents classification metrics for detecting elevated core temperature episodes using a fixed threshold of 38.0°C , including the area under the ROC curve (AUC), sensitivity, specificity.

Across most of the metrics, the proposed RCAF model achieves better performance than the baseline methods. It achieves the lowest RMSE(0.31°C) and MAE(0.23°C), the highest R^2 (0.67), and the lowest SD (0.31°C), with an MD of -0.05°C . RCAF also shows stronger ability to identify high-risk core temperature episodes, with the highest AUC (0.9038) and sensitivity (0.8540), confirming its reliability in safety-critical classification performance.

Table 6.1: Comprehensive evaluation metrics for all models on the full dataset.

Method	RMSE (°C)	MAE (°C)	R^2	MD $\pm$ SD (°C)
EKF	0.39	0.30	0.48	-0.07 ± 0.39
ECTemp-S	0.40	0.29	0.47	-0.06 ± 0.39
BKFB	0.48	0.37	0.24	-0.10 ± 0.47
MAKF	0.38	0.29	0.52	-0.02 ± 0.38
LSTM	0.45	0.35	0.31	0.00 ± 0.45
RCAKF	0.31	0.23	0.67	-0.05 ± 0.31

Method	AUC	Sensitivity	Specificity
EKF	0.8949	0.8116	0.8279
ECTemp-S	0.8950	0.8188	0.8221
BKFB	0.8858	0.8180	0.7852
MAKF	0.8963	0.8380	0.7992
LSTM	0.7563	0.5812	0.8414
RCAKF	0.9038	0.8540	0.7978

6.4 Discussion

This study proposes a Residual-Compensated Adaptive Kalman Filter (RCAKF) for continuous core temperature estimation using heart rate as the sole input signal. Unlike traditional filtering methods that rely on fixed-form observation models, the RCAKF integrates a lightweight LSTM module to dynamically compensate for structured residuals in the heart rate–temperature relationship. In addition, it incorporates an adaptive observation noise estimation mechanism to account for nonstationary physiological responses across diverse thermal and activity conditions. Together, these enhancements enable more robust and personalized temperature tracking across unseen experimental trials.

Visualization of the heart rate–core temperature mapping before and after residual compensation (Figure 6.2) reveals systematic, state-dependent deviations in the observation model. Although the residuals $r_t = HR_t - \hat{H}\hat{R}_t$ from the uncompensated EKF model are approximately zero-mean overall, their distribution deviates markedly from the Gaussian assumption. During *Exercise*, heart rate is consistently underestimated, while during *Rest* and *Recovery*, it is often overestimated. This asymmetric pattern violates the EKF’s assumption that observation noise is independent, identically distributed, and Gaussian.

These deviations not only challenge the statistical assumptions underpinning the EKF, but also degrade its ability to track physiological dynamics under varying conditions. The fixed-form observation model lacks the flexibility to capture the nonlinear, state-specific relationship between heart rate and core temperature, highlighting the need for a more adaptive approach. Crucially, the structured, time-varying nature of these residuals demands a correction mechanism capable of modeling temporal dependencies. Long Short-Term Memory (LSTM) networks are particularly well-suited for this task, as their gated architecture enables the modeling of both short- and long-term temporal dependencies [240]. This capability makes them highly effective in capturing the dynamic and state-dependent patterns characteristic of physiological time series such as heart rate.

To address these limitations, the proposed RCAF introduces two key innovations to enhance robustness and adaptability. First, an LSTM-based residual correction module learns time-dependent, state-specific deviations not captured by the static polynomial observation model. This correction is applied additively to the predicted heart rate, enabling the filter to account for structured residual patterns that would otherwise bias estimation. As a result, the corrected observation values exhibit more symmetric, zero-centered residual distributions across all physiological states, helping to restore compatibility with the Gaussian noise assumption required by the Kalman framework. Second, RCAF incorporates an adaptive noise estimation mechanism that dynamically updates the observation noise variance using a sliding window of recent prediction errors. This allows the filter to track fluctuations in uncertainty over time, particularly during transitions between physiological states or under measurement perturbations.

Importantly, these enhancements enable RCAF to maintain accurate and stable performance under nonstationary conditions, while preserving the interpretability and recursive structure of classical extended Kalman filters. The latent core temperature state remains explicitly estimated by the Kalman filter, and the LSTM’s influence is strictly localized to predicting additive corrections for the observation model. To avoid the complexity and instability associated with differentiable Kalman filter formulations, the LSTM module is trained separately using supervised learning to

minimize the error between observed and predicted heart rate. During inference, its output is used as a fixed additive term in the observation equation.

This design choice means that the entire RCAF framework is not trained in an end-to-end manner: gradients do not propagate through the Kalman filter equations, nor is the filter unrolled during training. Instead, the filter operates in a standard recursive manner, using the output of the pre-trained LSTM as a plug-in correction term. This sidesteps the need for implementing differentiable Kalman updates—often a nontrivial task—and ensures numerical stability, implementation simplicity, and clearer interpretability of each module’s role. It is highlighted to avoid confusion about the training pipeline and to emphasize that RCAF remains a hybrid system with cleanly decoupled training and inference phases.

Notably, the LSTM module was trained using only heart rate, without requiring contextual labels or environmental features. This minimalist input design enhances the model’s transferability to unseen scenarios and reduces reliance on high-dimensional sensing data, which is advantageous for real-world deployment in wearable systems.

To evaluate the performance of the proposed RCAF, it was compared against five representative models: the standard extended Kalman filter (EKF), the improved ECTemp model with a sigmoid observation function (ECTemp-S), the biphasic Kalman filter-based model (BKFB), the moving average Kalman filter (MAKF), and a standalone LSTM network.

The EKF serves as a standard and widely used approach for physiological state estimation [173]. It applies fixed-form state transition and observation models and assumes observation noise to be independent and identically distributed Gaussian. While it achieves moderate performance ($\text{RMSE} = 0.39$, $R^2 = 0.48$, $\text{AUC} = 0.895$), it cannot account for structured, state-dependent residuals and thus lacks adaptability under dynamic physiological conditions.

The ECTemp-S model modifies the standard EKF by replacing its quadratic observation function with a generalized logistic (sigmoid) mapping from core temperature to heart rate. This adjustment is intended to better reflect the saturating relationship between heart rate and temperature at physiological extremes. In

practice, however, ECTemp-S yields performance that is largely comparable to EKF across regression and classification metrics (RMSE = 0.40, $R^2 = 0.47$, AUC = 0.895). Slight improvements in sensitivity (0.819) and marginally reduced bias (MD $\pm$ SD = $-0.06 \pm 0.39^\circ\text{C}$) were observed, suggesting that while the sigmoid formulation aligns better with physiological trends, its practical impact under the tested conditions is limited.

The MAKF model—introduced under this name in the current study—extends the EKF by incorporating a phase-switching mechanism based on moving-average heart rate trends. This enables it to differentiate between *exercise* and *recovery* phases, resulting in comparable regression performance but slightly improved classification accuracy over EKF (RMSE = 0.38, $R^2 = 0.52$, AUC = 0.896), while maintaining a relatively low bias (MD $\pm$ SD = $-0.02 \pm 0.38^\circ\text{C}$).

The BKFB method, adapted from Falcone et al. [148], segments the estimation into two fixed filters triggered by a threshold on heart rate. Although conceptually simple, this binary switching strategy leads to degraded performance across both regression and classification metrics (RMSE = 0.48, $R^2 = 0.24$, AUC = 0.886), indicating limited generalizability.

The standalone LSTM model, which directly regresses core temperature from a 10-minute heart rate window, yields the weakest performance overall (RMSE = 0.45, $R^2 = 0.31$), with low sensitivity (0.581) despite moderate specificity (0.841). Its purely data-driven design lacks the recursive estimation and uncertainty modeling capabilities of Kalman-based methods.

In contrast, RCAF consistently achieves better performance than all baselines across almost all of the evaluation metrics. It achieves the lowest error (RMSE = 0.31, MAE = 0.23), highest explanatory power ($R^2 = 0.67$), and best residual distribution (MD $\pm$ SD = $-0.05 \pm 0.31^\circ\text{C}$). In the classification task at the 38.0°C threshold, it maintains strong discriminative performance (AUC = 0.904) and the highest sensitivity (0.854). This elevated sensitivity is particularly critical in safety-sensitive applications, where the timely detection of elevated core temperature is essential for preventing heat-related illnesses. While the improvement in AUC over certain baselines is relatively modest, the substantial reduction in RMSE is of greater

practical significance. A more accurate estimation of the degree of hyperthermia allows for finer-grained monitoring and better-informed interventions, which are crucial in real-world deployment scenarios.

These results highlight the strength of RCAF’s hybrid architecture, which combines two complementary mechanisms: a learnable LSTM-based residual correction module that accounts for structured, state-dependent deviations in the observation model, and an adaptive observation noise estimator that dynamically adjusts uncertainty based on recent prediction errors. Together, these components enable RCAF to flexibly adapt to complex physiological variations while maintaining the interpretability, stability, and real-time compatibility of classical Kalman filtering frameworks.

Contextualizing the performance of the RCAF model is important, though direct comparisons with other method types are challenging due to differences in protocols and reported metrics. Nevertheless, a qualitative assessment can be made against established benchmarks. For example, specialized non-invasive methods like zero-heat-flux (ZHF) thermometry can achieve very high precision, with reported 95% limits of agreement of $\pm 0.40^{\circ}\text{C}$ against esophageal temperature in controlled exercise studies [39]. The RCAF model’s 95% limits of agreement of $\pm 0.61^{\circ}\text{C}$, while not reaching this level of precision, are a strong result for an estimation framework that relies only on a single, universally available physiological signal. The primary contribution of this approach, therefore, is not in surpassing the accuracy of hardware-intensive methods, but in offering a compelling balance of acceptable performance, convenience, and scalability for real-world, continuous monitoring where specialized equipment is impractical.

While RCAF improves estimation performance, several limitations of this study highlight important avenues for future research.

First, the model was validated on a cohort of young, healthy adults, and its generalizability to other populations warrants discussion. Thermoregulatory capabilities, for instance, vary significantly across different populations. Children have underdeveloped cardiovascular and sweating functions, while older adults often exhibit diminished thermoregulatory responses. In addition, age-related changes in

cardiovascular regulation may result in less predictable heart rate and blood pressure responses during thermal and physical stress [241]. These age-related physiological differences would likely alter the fundamental relationship between heart rate and core temperature, requiring age-specific model calibration to ensure accuracy [242]. Similarly, individuals with underlying health conditions (e.g., cardiovascular disease) or those on medications that affect heart rate or blood pressure may not fit the model’s assumptions [243]. Future work should aim to validate and adapt the model on more diverse populations to broaden its applicability.

Second, the residual-compensation LSTM component was trained on data featuring relatively homogeneous activity profiles, consisting mainly of structured transitions between rest, exercise, and recovery. As such, the LSTM may partially overfit to these stereotyped temporal patterns, capturing activity transitions rather than more general physiological dynamics. Future work should therefore investigate the model’s robustness using datasets that include more diverse and unstructured activity types to better assess its generalizability.

Third, the model’s real-world performance is linked to the quality of the input heart rate signal. The proposed RCAF, with its adaptive filtering mechanism, is inherently designed to have robustness against noisy data. Furthermore, modern commercial heart rate sensors have demonstrated substantial improvements in accuracy, allowing them to provide reasonably reliable measurements even under exercise conditions [244, 245]. However, the potential impact of extreme noise events or artifacts on the model’s estimation accuracy remains an area that warrants further investigation. Future work could explore the model’s performance boundaries under such challenging conditions.

Finally, a more profound limitation concerns the model’s specialization in hyperthermic states. The evaluation dataset exclusively captures physiological responses to exertional heat stress, meaning the RCAF model is tuned for the dynamics of a warming body but is naive to the complexities of cooling. This is a critical point because the body’s thermoregulatory responses to cold are fundamentally different from those to heat. In a hot environment or during exercise, heart rate reliably increases to meet elevated metabolic demands and to help dissipate heat.

In a cold environment, however, the body employs two different primary strategies: peripheral vasoconstriction to reduce heat loss from the skin, and an increase in metabolic heat production (e.g., shivering) to generate warmth [246, 247]. These distinct physiological mechanisms create a more complex relationship between heart rate and core temperature that is not captured in the training data. Therefore, a model trained solely on heat-stress data cannot be expected to generalize accurately to hypothermic conditions. Future work should focus on developing a more robust framework, potentially a state-switching model, capable of identifying and adapting to these different physiological states to ensure accuracy across a wider range of thermal environments.

Despite these limitations, the proposed RCAF framework offers a practical advantage by relying solely on heart rate as the input signal. This design choice provides several benefits. Heart rate is readily available from most commercial wearable devices with minimal setup and calibration [210, 248]. By avoiding dependence on additional sensor modalities, RCAF minimizes hardware complexity, computational overhead, and power consumption—making it well-suited for deployment on lightweight, low-power wearable platforms. In contrast, multimodal models that incorporate skin temperature or heat flux often require bulky or intrusive sensors, which may reduce user comfort and long-term wearability. Moreover, skin temperature is known to exhibit substantial variability depending on sensor placement and ambient conditions, and it often poorly correlates with core body temperature during exertional heat stress [35, 37]. These factors introduce inconsistencies and may limit the generalizability of multimodal approaches in diverse environments.

Future work could explore several promising directions to enhance the RCAF framework. First, the selective integration of auxiliary signals such as accelerometry, sweat rate, or body surface temperature could improve the model’s context-awareness without compromising wearability. For example, recent studies on other measurement sites, like the oral cavity, highlight the importance of compensating for environmental disturbances, a task for which auxiliary sensors would be highly valuable [249]. Second, this contextual data could be used to enhance the residual correction module, allowing it to dynamically adapt based on detected activity states or

environmental conditions, which would further improve performance in nonstationary settings. Finally, incorporating uncertainty estimation within the LSTM component itself would be a significant step forward, enabling the RCAKF to provide not just an estimate but also a confidence level for each prediction, which is crucial for safety-critical applications.

6.5 Conclusion

This study presented a Residual-Compensated Adaptive Kalman Filter (RCAKF), a novel hybrid framework for real-time core body temperature estimation using only a heart rate signal. The proposed method enhances the classical extended Kalman filter by integrating an LSTM-based residual correction module and an adaptive observation noise estimator. Experimental results on a controlled physiological dataset demonstrated the strength of the RCAKF. It consistently achieved better performance than five representative models across both regression and classification tasks. Specifically, the RCAKF achieved a Root Mean Squared Error (RMSE) of 0.31°C and a coefficient of determination (R^2) of 0.67, the best performance among all tested methods. It also showed the highest sensitivity in detecting elevated core temperature episodes, which is critical for safety monitoring. These findings highlight the effectiveness of combining model-based filtering with lightweight machine learning augmentation to achieve robust and deployable thermal monitoring. Future work will focus on validating the RCAKF in free-living settings and exploring its extension with uncertainty-aware decision support for personalized heat stress management.

7

A Physics-Guided Hybrid Model for Core Body Temperature Estimation: Thermoregulation model with Residual Learning

Chapter 6 demonstrated that integrating deep learning with a Kalman filter effectively mitigates estimation errors through data-driven residual compensation. However, while effective, that approach relies on a statistical observation model lacking explicit physiological constraints, which limits interpretability and robustness when data quality degrades.

To overcome the limitations of purely data-driven approaches, this chapter presents a physics-guided hybrid model for core temperature estimation. A modified two-node thermoregulation model compatible with wearable sensor data is used to generate a baseline physiological trajectory. Then the baseline is coupled with a light-weight temporal neural network (convolutional neural network (1D-CNN)/long short-term memory (LSTM)/gated recurrent unit (GRU)), which compensates for the residual between the first-principles model prediction and the true core temperature.

Evaluated on a controlled dataset of 22 participants (the same Dataset 1 used in Chapter 5 and Chapter 6), the hybrid framework with CNN achieved the lowest estimation error ($\text{RMSE} = 0.27\text{ }^{\circ}\text{C}$), better than physics-only models, standard EKFs, and purely data-driven neural networks. Notably, the hybrid framework with the LSTM variant demonstrated higher resilience to extreme heart rate noise.

These findings demonstrate that embedding physiological priors within a learning framework can provide a satisfying balance between accuracy, robustness,

and interpretability.

7.1 Introduction

Exertional heat stroke (EHS) is one of the most life-threatening outcomes of thermal dysregulation. EHS frequently occurs in athletes, military personnel, and outdoor laborers operating under high-heat and high-humidity conditions. The mortality rate of EHS can exceed 25% if timely intervention is not provided [250, 251]. In the United States, heat-related illnesses lead to approximately 4,100 emergency department visits each year, with over half (54.6%) resulting in hospitalization and 3.5% ending in death either in the emergency department or during hospitalization [252]. Because elevated CBT is the defining feature of heat stroke, early detection and continuous monitoring are critical for preventing irreversible damage or death.

Gold-standard methods for CBT measurement—such as rectal, esophageal, and ingestible thermometry—while highly accurate, are generally unsuitable for continuous, long-term, or field-based applications due to their invasiveness, discomfort, and logistical constraints. As a more practical alternative, temperature measurements are often taken from peripheral sites such as the axilla, tympanic membrane, oral cavity, and forehead, using either contact thermometers or non-contact infrared thermography (IRT) [31, 32, 35]. Despite their convenience, these approaches estimate superficial temperatures that frequently diverge from true CBT, particularly during physical exertion or environmental heat stress. This discrepancy arises from peripheral vasomotor activity, evaporative cooling, and external temperature influences, with deviations from actual CBT exceeding 2–3°C in many cases [36, 37, 106].

Beyond direct measurement approaches, physiology-based thermal models have long been developed to simulate CBT by capturing the underlying mechanisms of human heat exchange and thermoregulation. These models typically consist of two interacting components: a passive system governed by physical heat transfer laws—including conduction, convection, radiation, and perfusion—and an active system driven by physiological responses such as vasomotion, sweating, and shivering [196, 253]. The foundational basis for many of these models is the Pennes

bioheat equation, which describes internal heat generation and redistribution through blood flow [254].

A widely adopted formulation is the two-node model, originally introduced by Gagge, which simplifies the body into two compartments: core and skin [255, 256]. While computationally efficient, this approach offers limited spatial resolution. To enhance anatomical realism and model localized thermal responses, multi-node and multi-layer models have been developed by researchers such as Stolwijk and Fiala [47, 257]. These advanced models divide the body into multiple segments and concentric tissue layers (e.g., skin, fat, muscle, bone), enabling more precise simulation of both heat transport and thermoregulatory dynamics under varied environmental and metabolic conditions.

These models are valued for their strong physiological interpretability and their compatibility with environmental and clothing simulation frameworks. However, their practical use in real-time or individualized applications remains constrained. Accurate predictions often require detailed personal parameters—such as metabolic rate, body composition, and surface area—and the models can be highly sensitive to parameter tuning. Moreover, they typically depend on pre-defined parameter settings or boundary conditions established prior to simulation, making them difficult to deploy as real-time adaptive systems. In addition, fixed structural assumptions and high computational complexity further hinder their adaptability to inter-individual variability and dynamically changing physiological states [199, 200].

To address the limitations of traditional physiological models, recent research has increasingly focused on data-driven approaches to estimate CBT. Leveraging physiological signals that are readily obtainable from wearable sensors—such as heart rate (HR), skin temperature, and heat flux—machine learning models have been developed to learn empirical mappings between these input signals and CBT. These models offer several advantages: they require fewer assumptions about human anatomy and thermoregulation, can be rapidly adapted to different individuals and conditions, and are capable of capturing nonlinear relationships that may be difficult to model mechanistically [146, 147].

Techniques explored in prior work include linear and nonlinear regression methods [49, 150], as well as deep learning architectures like convolutional and recurrent neural networks [149, 175, 258]. These approaches have shown promising performance in laboratory and semi-free-living environments, especially when trained on sufficient quantities of high-quality, labeled data.

In parallel, Kalman filtering methods based on data-driven state-space models have also been widely employed for CBT estimation. These methods treat CBT as a latent dynamic state and use physiological signals such as heart rate or skin temperature as observations to iteratively update the estimate over time. The state transition and observation models are typically parameterized either using empirical data or trained directly on a designated dataset [38, 46, 138]. This approach allows the filter to capture individual-specific thermal dynamics and adapt to varying physiological and environmental conditions, while preserving the advantages of Kalman filtering, including temporal consistency and robustness to measurement noise.

However, data-driven models also come with inherent challenges. They often act as “black boxes,” providing limited physiological interpretability, and may generalize poorly when deployed across diverse populations, environmental conditions, or unseen activities. These limitations have motivated the development of hybrid modeling strategies that combine the interpretability of physics-based models with the adaptability of machine learning [50, 259].

In this context, hybrid modeling approaches have emerged as a promising direction to overcome the respective limitations of purely physics-based and purely data-driven methods. These approaches seek to embed physiological knowledge into learning frameworks, thereby enhancing model interpretability while retaining the flexibility to learn complex residual patterns from data.

In this study, an improved physics-guided hybrid model for core body temperature (CBT) estimation is proposed by integrating a modified two-node thermoregulation model with a neural-network-based residual correction framework. Physics-informed machine learning (PIML) refers to approaches that incorporate established physical or physiological knowledge into data-driven models, rather than relying solely on

statistical relationships learned from data [260]. In the proposed framework, the two-node model provides a physiologically interpretable baseline estimate based on human heat-balance principles, while the neural network learns the remaining discrepancy between the model-derived estimate and the measured CBT. In this way, the physical model constrains the overall estimation process, whereas the data-driven component compensates for modelling simplifications, parameter uncertainty, and individual variability.

To the best of our knowledge, this is the first study to extend the concept of physics-informed machine learning (PIML) to wearable CBT estimation. While PIML has been widely applied in fields such as fluid dynamics, structural mechanics, and biomedical imaging, its use in human thermoregulation and physiological monitoring has not been previously explored [260–262].

It is important to note that the present framework adopts a residual modeling paradigm within the broader PIML definition, in which a physics-based thermoregulation model generates the baseline prediction, and a neural network is trained to model the residual discrepancy between the simulated and measured CBT. This formulation differs fundamentally from the commonly used Physics-Informed Neural Network (PINN) approach, where governing differential equations are embedded directly into the network loss function as soft constraints. The residual-learning paradigm retains an explicit physical model that operates independently and is adaptively corrected by data-driven learning, thereby preserving physiological interpretability while enhancing flexibility and robustness to model uncertainty. In this formulation, heart rate dynamically estimates the metabolic rate (M), and the measured skin temperature substitutes for the modeled skin node, enabling real-time physiological coupling between the core and the environment. The two-node model then estimates the dynamic evolution of CBT based on inter-compartmental heat exchange and sensor-derived inputs. To further account for unmodeled physiological and environmental influences, the neural network learns the residual error between the physical-model predictions and the measured CBT. This hybrid design thus bridges interpretable thermal regulation with adaptive neural learning for wearable, real-time CBT monitoring.

To validate the proposed framework, three hybrid variants integrating the two-node thermoregulation model with different neural network architectures are evaluated using multi-subject data collected under diverse thermal and activity conditions: one-dimensional convolutional neural network (1D-CNN), long short-term memory (LSTM), and gated recurrent unit (GRU). The hybrid models are compared against the standalone physics-based formulation, the widely used extended Kalman filter (EKF), and purely data-driven neural models to comprehensively assess predictive accuracy and robustness. Results indicate that the 1D-CNN-based hybrid performs best under low-to-moderate heart rate noise, while the LSTM-based hybrid maintains stable performance even under high noise levels, demonstrating complementary robustness across conditions. Moreover, the proposed models are computationally lightweight, requiring only a few hundred thousand floating-point operations per inference, making them suitable for real-time execution on wearable or edge devices. This physics-guided hybrid design provides a physiologically meaningful decomposition between modeled and learned components, offering a path toward interpretable, generalizable, and on-device CBT monitoring.

7.2 Methods

7.2.1 Two-Node Thermoregulation Model

A physiologically grounded two-node thermoregulation model is employed *and adapted* to describe human heat-exchange dynamics under varying environmental and metabolic conditions. Originally proposed by Gagge et al. and widely adopted in thermal comfort and physiology research, the model partitions the human body into two distinct thermal compartments: a core region representing vital organs and deep tissues, and a skin region encompassing peripheral tissues and the body surface [255, 256].

A schematic overview of the canonical model structure is shown in Figure 7.1. In the classical formulation, internal heat is primarily generated in the core via metabolic activity, transferred to the skin through blood perfusion and tissue conduction, and eventually dissipated to the ambient environment through convective, radiative, and evaporative heat flows ($q_{sk \rightarrow am}$). Thermoregulatory responses such as vasomotor

control and sweating are incorporated as nonlinear feedback mechanisms that modulate these heat-transfer rates based on deviations from physiological set-points. In the present work, a measurement-driven variant of this framework is adopted, in which wearable-measured skin temperature (T_{sk}) replaces the computed skin-node temperature, thereby implicitly capturing the combined convective, radiative, and evaporative effects represented by $q_{sk \rightarrow am}$. This substitution is indicated by a dashed connection between the skin node and the ambient environment in Figure 7.1, representing that the skin–ambient heat-loss pathway is no longer explicitly solved but constrained by real sensor measurements. Meanwhile, heart rate (HR) is used to dynamically estimate metabolic heat production (M), enabling real-time, physiology-driven adaptation of the model.

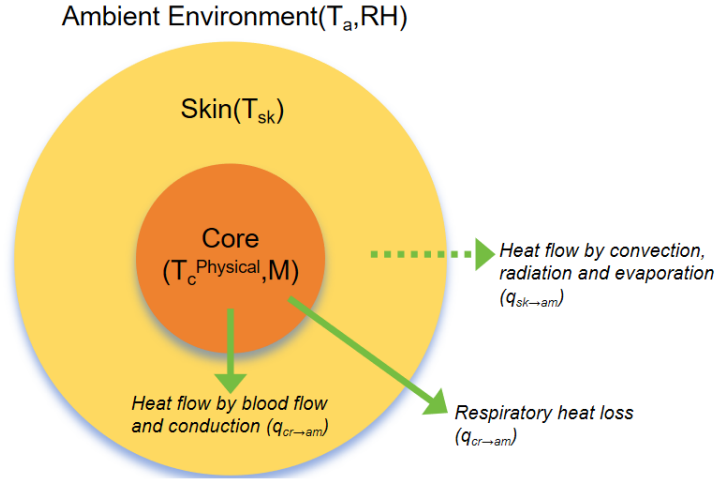


Figure 7.1: Schematic of the two-node thermoregulation model. Heat is generated in the core through metabolic processes and transferred to the skin via blood flow and tissue conduction. The core also loses heat directly to the environment through respiratory heat exchange. The skin compartment exchanges heat with the environment through convection, radiation, and evaporation.

The dynamic behaviour of the CBT is described by a first-order energy balance equation, which accounts for internal heat production, heat transfer to the skin, and direct heat loss to the ambient environment through respiration. The governing equation is:

$$\frac{dT_c}{dt} = \frac{A_{body}}{m_{cr} C_{cr}} (M - W - q_{cr \rightarrow sk} - q_{cr \rightarrow am}) \quad (7.2.1.1)$$

In this equation, A_{body} denotes the body surface area (m^2), m_{cr} is the mass of the core compartment (kg), which is defined by a fixed proportion of total body mass as $m_{cr} = (1 - \alpha) \cdot m_{body}$, with $\alpha = 0.042$. C_{cr} is its specific heat capacity (J/kg/°C). The term M represents metabolic heat production per unit body surface area (W/ m^2), which is estimated from heart rate measurements as described in Equation 7.2.1.6. External mechanical work W is assumed to be zero in this study, as the majority (typically over 75–80%) of the metabolic energy produced during physical activity is converted into heat rather than external work [263]. The term $q_{cr \rightarrow sk}$ denotes the heat flux from the core to the skin compartment, while $q_{cr \rightarrow am}$ accounts for respiratory heat loss from the core to the ambient environment.

The heat flux from the core to the skin compartment, $q_{cr \rightarrow sk}$, represents the combined effect of convective heat transfer via blood perfusion and conductive heat flow through tissues. It is modeled as:

$$q_{cr \rightarrow sk} = \left(c_{bl} \rho_{bl} \cdot \frac{V_{bl}}{3600 \cdot 1000} + K_{\min} \right) (T_c - T_{sk}) \quad (7.2.1.2)$$

In this expression, c_{bl} and ρ_{bl} are the specific heat capacity (J/kg/°C) and density (kg/ m^3) of blood, respectively. The coefficient $K_{\min}$ represents the minimum thermal conductance between the core and skin tissues (W/ m^2 /K). The temperatures T_c and T_{sk} are the instantaneous core and skin temperatures, respectively.

V_{bl} denotes the skin blood flow rate per unit body surface area (L/ m^2 /h). It is dynamically regulated in response to thermoregulatory signals. It increases with rising CBT (warm signal) and decreases when skin temperature falls below its set-point (cold signal). This regulatory mechanism is modeled as:

$$V_{bl} = V_{bl, \text{basal}} + \frac{C_{bl, \text{dil}} \cdot \max(0, T_c - T_{c, \text{set}})}{1 + C_{bl, \text{str}} \cdot \max(0, T_{sk, \text{set}} - T_{sk})} \quad (7.2.1.3)$$

Here, $C_{bl, \text{dil}}$ and $C_{bl, \text{str}}$ are the vasodilation and vasoconstriction coefficients, respectively. $T_{c, \text{set}}$ and $T_{sk, \text{set}}$ are the physiological set-point temperatures for the core and skin. The formulation ensures that blood flow increases rapidly when CBT exceeds its set-point, but this effect is attenuated under peripheral vasoconstriction conditions when skin temperature falls below its set-point.

The core also loses heat directly to the ambient environment through respiration, which includes both sensible and latent heat exchange via the inhalation and exhalation of air. The respiratory heat loss $q_{cr \rightarrow am}$ is modeled as:

$$q_{cr \rightarrow am} = 1.4 \times 10^{-3} \cdot M \cdot (34 - T_a) + 0.0023 \cdot M \cdot (44 - 0.0075 \cdot P_a) \quad (7.2.1.4)$$

In this equation, M is the metabolic heat production per unit body surface area (W/m^2), T_a is the ambient air temperature ($^{\circ}\text{C}$), and P_a is the ambient water vapor pressure (Pa). The first term represents sensible heat loss due to the temperature gradient between exhaled air and the environment, while the second term represents latent heat loss due to the difference in water vapor pressure between saturated exhaled air and ambient air.

The ambient water vapor pressure P_a is estimated based on the relative humidity (RH) and ambient air temperature T_a using the Tetens formula for saturation vapor pressure [264]. The calculation proceeds as:

$$P_a = RH \cdot 0.61078 \cdot \exp\left(\frac{17.27 \cdot T_a}{T_a + 237.3}\right) \cdot 1000 \quad (7.2.1.5)$$

Here, RH is the relative humidity expressed as a fraction between 0 and 1.

A simplified version of the respiratory heat loss model can be obtained by setting $T_a = 25^{\circ}\text{C}$ and replacing P_a with a constant reference value $\bar{P}_a = 1584$ Pa, corresponding to the vapor pressure at 25°C , 50% RH. This simplified form eliminates the need for real-time environmental measurements, thereby reducing model dependency on environmental sensors. It allows the model to operate using only heart rate and forehead skin temperature as inputs. The complete and simplified forms can be selectively applied depending on the availability of ambient humidity and temperature measurements.

The metabolic heat production rate M is estimated from heart rate measurements [265]. The formula relates heart rate to metabolic rate through the subject's age and body mass, assuming standardized basal conditions:

$$M = M_0 + \left(\frac{MWC/A_{body} - M_0}{HR_{\max} - HR_0}\right) (HR - HR_0) \quad (7.2.1.6)$$

Here, HR is the measured heart rate (beats per minute), HR_0 is the resting heart rate, and $HR_{\max} = 208 - 0.7 \cdot \text{age}$ is the estimated maximum heart rate. The maximal working capacity MWC (W) is computed as:

$$MWC = (19.45 - 0.133 \cdot \text{age}) \cdot m_{\text{body}} \quad (7.2.1.7)$$

where m_{body} is the assumed body mass. The resulting M is normalized by body surface area to yield the metabolic rate per unit area (W/m^2), which is used in the CBT equation (7.2.1.1).

This approach allows indirect estimation of individual thermal load in field or wearable settings without requiring direct calorimetry or gas analysis.

All constants used in the model, including thermophysical properties, regulatory coefficients, and physiological assumptions, are summarized in Table 7.1. These values are primarily drawn from established literature on human thermoregulation models [255, 256, 265].

Table 7.1: Summary of model constants used in the two-node thermoregulation model

Symbol	Description	Value	Unit
C_{cr}	Specific heat capacity of core tissue	3650	J/kg/°C
c_{bl}	Specific heat capacity of blood	4180	J/kg/°C
ρ_{bl}	Density of blood	1050	kg/m ³
$K_{\min}$	Minimum tissue thermal conductance	5.28	W/m ² /K
$T_{c,\text{set}}$	CBT set-point	36.8	°C
$T_{sk,\text{set}}$	Skin temperature set-point	33.7	°C
$V_{bl,\text{basal}}$	Basal skin blood flow	6.3	L/m ² /h
$C_{bl,\text{dil}}$	Vasodilation coefficient	200	L/m ² /h/°C
$C_{bl,\text{str}}$	Vasoconstriction sensitivity	0.1	1/°C
A_{body}	Body surface area	1.8	m ²
m_{body}	Total body mass	70	kg
age	Age	30	years
α	Skin mass ratio	0.042	—
HR_0	Resting heart rate	60	bpm
M_0	Resting metabolic rate	60	W/m ²

7.2.2 Residual Compensation with Neural Networks

The modified two-node model provides first-principles predictions of CBT, yet residual errors remain due to simplified assumptions, inter-individual variability, and sensor noise. A lightweight neural network-based residual learner is therefore introduced to refine the physics-based prediction.

At time t , the residual target is defined as

$$\varepsilon_t = T_{c,\text{true}}(t) - T_{c,\text{physical}}(t), \quad (7.2.2.1)$$

where $T_{c,\text{physical}}(t)$ is produced by the two-node model. The network learns a mapping

$$\hat{\varepsilon}_t = f_{\theta}(\mathbf{x}_{t-k+1:t}),$$

from a length- k input sequence $\mathbf{x}_{t-k+1:t}$ to the residual. The final compensated estimate is

$$T_{\text{core,final}}(t) = T_{c,\text{physical}}(t) + \hat{\varepsilon}_t. \quad (7.2.2.2)$$

A hybrid input configuration aligned with the two-node model is adopted:

$$\mathbf{x}_t = [HR, T_{sk}, T_a, RH, T_{c,\text{physical}}]_t,$$

where HR is heart rate, T_{sk} is wearable skin temperature (head), T_a ambient temperature, RH relative humidity, $T_{c,\text{physical}}$ the physics output. All features and residuals are min-max scaled using statistics fit on the training data. For each participant, sliding windows $\mathbf{x}_{t-k+1:t}$ are constructed, where the window length k is selected via grid search on the validation set ($k \in \{5, 10, 15, 20\}$).

To implement the residual learner $f_{\theta}(\cdot)$, three lightweight neural architectures were examined: LSTM, GRU, and 1D-CNN [266–268]. These represent three common paradigms for modeling temporal physiological signals—sequential recurrence, gated memory, and local temporal convolution, respectively. To ensure a fair comparison across architectures, all models were designed with comparable parameter counts, network depth, and input configurations, so that any performance difference primarily reflects their structural characteristics rather than differences in model size or parameter quantity. The architectures were also deliberately kept small to enable potential deployment on wearable or mobile devices with limited computational resources.

The detailed network configurations, training hyperparameters, and computational characteristics are summarized in Table 7.2. A schematic illustration of the integrated hybrid framework, including the two-node physical model and the residual compensation module, is provided in Figure 7.2. For clarity, the overall online estimation procedure is summarized in Algorithm 3, where the physics-based model

operates in an open-loop manner and the neural residual learner is applied in real time to refine its output. To quantify deployment cost, all FLOPs and latency numbers reported in Table 7.2 were profiled on a consumer laptop with an AMD Ryzen 7 processor (no GPU acceleration), using Python 3.10 and TensorFlow 2.20; wall-clock latency is the median over 100 runs and may vary across hardware.

Table 7.2: Neural residual learners: architectures, training settings, and computational characteristics.

Attribute	LSTM residual learner	GRU residual learner	1D-CNN residual learner
Structure	LSTM(64) Dropout(0.2) LSTM(32) Dropout(0.2) Dense(1)	GRU(80) Dropout(0.2) GRU(40) Dropout(0.2) Dense(1)	Conv1D(32, kernel=3) BatchNorm ReLU MaxPool Conv1D(64, kernel=3) BatchNorm ReLU MaxPool Conv1D(96, kernel=3) BatchNorm ReLU GlobalAvgPool Dense(32) Dense(1)
Input window k (samples)	10	10	15
Loss	Mean Squared Error (MSE)		
Optimizer	Adam		
Learning rate	0.001		
Batch size	32		
Epoch	200		
Early stopping patience	30		
Dropout rate	0.2		
Trainable parameters ($\times 10^3$)	~31	~32	~35
Size (KB)	~120	~140	~140
FLOPs per step ($\times 10^3$)	~600	~700	~250
Latency (ms / step)	~50	~50	~50

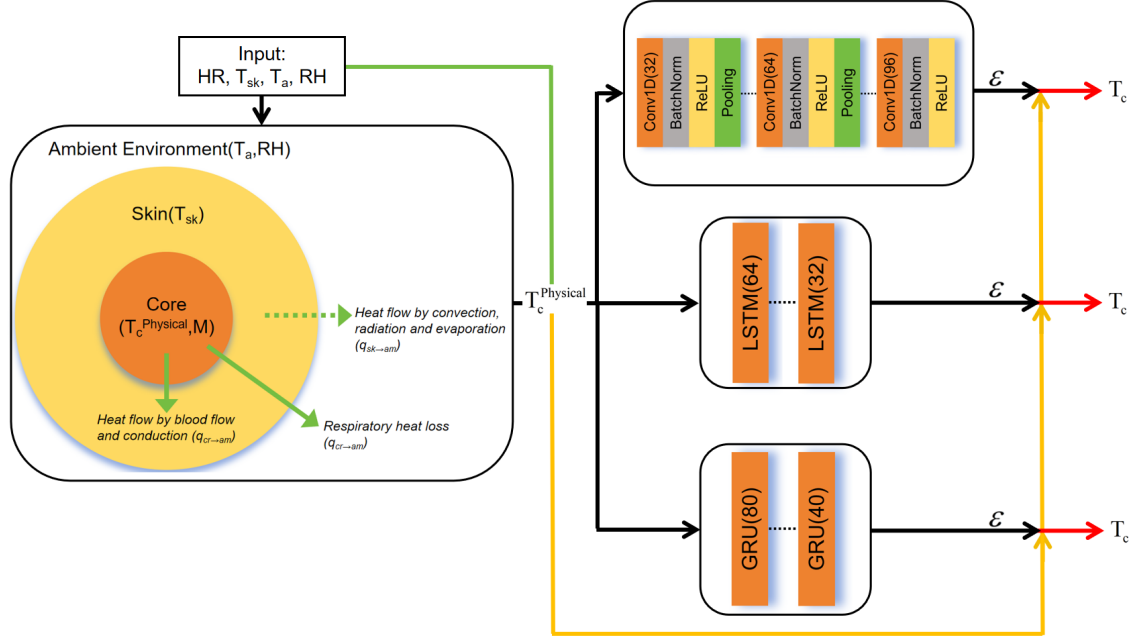


Figure 7.2: Overall architecture of the proposed physics-guided hybrid model: The two-node thermoregulation module provides a first-principles estimate of CBT, while the neural residual learner (LSTM, GRU, or 1D-CNN) refines the output by modeling residual temporal patterns. Wearable signals (heart rate and skin temperature) and environmental variables serve as inputs to both components.

Algorithm 3 Online Hybrid Core Temperature Estimation

Inputs: $HR_t, T_{sk,t}, T_{a,t}, RH_t$; sampling interval Δt ; trained residual learner $f_\theta(\cdot)$; two-node model parameters

Output: Estimated CBT $T_{c,\text{final}}(t)$

for each new time step t **do**

 (1) *Physics-based prediction*

 Estimate metabolic rate M_t from HR_t (Eq. 7.2.1.6); set $W_t \leftarrow 0$

 Compute $q_{cr \rightarrow sk}(t)$ using previous $T_{c,\text{physical}}(t-1)$ and measured $T_{sk,t}$

 Compute $q_{cr \rightarrow am}(t)$ from $T_{a,t}$ and RH_t (Eq. 7.2.1.4)

 Update CBT prediction:

$$T_{c,\text{physical}}(t) = T_{c,\text{physical}}(t-1) + \frac{A_{\text{body}}}{m_{cr} C_{cr}} (M_t - q_{cr \rightarrow sk}(t) - q_{cr \rightarrow am}(t)) \Delta t$$

 (2) *Residual correction*

 Form feature vector $\mathbf{x}_t = [HR_t, T_{sk,t}, T_{a,t}, RH_t, T_{c,\text{physical}}(t)]$

 Construct input window $\mathbf{x}_{t-k+1:t}$ using the latest k samples

 Predict residual $\hat{\varepsilon}_t = f_\theta(\mathbf{x}_{t-k+1:t})$

 Correct prediction: $T_{c,\text{final}}(t) = T_{c,\text{physical}}(t) + \hat{\varepsilon}_t$

end for

return $T_{c,\text{final}}(t)$ at each time step

7.2.3 Dataset

This study utilized a controlled laboratory dataset comprising 22 healthy adult participants who performed treadmill-based exercise under a range of thermal conditions [150, 151]. The same dataset was used as Dataset 1 in Chapter 5 and as the dataset used in Chapter 6. A detailed description of the participants, experimental protocol, and measurements is provided in Table 5.1. Heart rate values were obtained directly from the chest ECG measurement system.

The dataset was split strictly at the participant level: 10 participants (39 trials) were used for training, 5 participants (17 trials) for validation, and 7 participants (21 trials) for testing. This ensured that no subject contributed data to more than one subset. The participant-level partitioning was kept identical across all evaluated models to guarantee a fair and consistent comparison.

7.2.4 Evaluation Metrics and Comparative Settings

The performance of all models was evaluated in terms of both absolute accuracy and temporal agreement with the ground-truth CBT. Six complementary metrics were used to quantify prediction accuracy, bias, explanatory power, and alignment, including the root mean squared error (RMSE), the mean absolute error (MAE), the mean difference (MD), the standard deviation (SD) of MD, coefficient of determination (R^2), and the dynamic time warping (DTW) distance.

To assess temporal consistency between predicted and reference trajectories, the dynamic time warping (DTW) distance is defined as:

$$\text{DTW}(X, Y) = \min_{w \in \mathcal{W}} \sqrt{\sum_{(i,j) \in w} (X_i - Y_j)^2}, \quad (7.2.4.1)$$

where X and Y denote the predicted and reference temperature sequences, respectively, and $\mathcal{W}$ is the set of all possible monotonic warping paths. Lower DTW values indicate better alignment of temporal patterns. Finally, Bland–Altman analysis was applied to examine overall agreement and potential systematic bias across the full temperature range.

To ensure a fair comparison, all models were trained and evaluated using identical data splits, normalization procedures, and temporal windowing strategies as described

in Section 7.2.2 and Section 7.2.3. Each model was optimized based on validation RMSE and evaluated on test set.

The following models were compared:

1. Physics-only model: the baseline two-node thermoregulation model without residual correction.
2. EKF model: an Extended Kalman Filter using the same inputs (HR and T_{sk}), adopting a first-order process model and a second-order polynomial observation model. The model parameters were trained from the training set [137, 173].
3. Hybrid-LSTM: the proposed physics-guided hybrid model with residual compensation using a two-layer Long Short-Term Memory network.
4. Hybrid-GRU: the hybrid model using a two-layer Gated Recurrent Unit network for residual learning.
5. Hybrid-CNN (full input): the hybrid model using a three-layer one-dimensional Convolutional Neural Network for residual learning, with the full feature vector $[HR, T_{sk}, T_a, RH, T_{c,physical}]$.
6. Data-driven neural models: the purely data-driven counterparts of the above hybrid architectures, including LSTM-only, GRU-only, and CNN-only networks, trained with the same input features but without coupling to the two-node physical model. These models directly map the observed signals to CBT.
7. Hybrid-CNN (simplified input): a lightweight variant using only HR and T_{sk} as inputs. This model adopts the simplified two-node thermoregulation formulation described in Section 7.2.1, where the respiratory heat loss component is simplified by fixing the ambient temperature to $T_a = 25^\circ\text{C}$ and substituting the ambient vapor pressure P_a with a constant reference value $\bar{P}_a = 1584$ Pa. This reduces dependency on environmental sensors while maintaining real-time applicability.

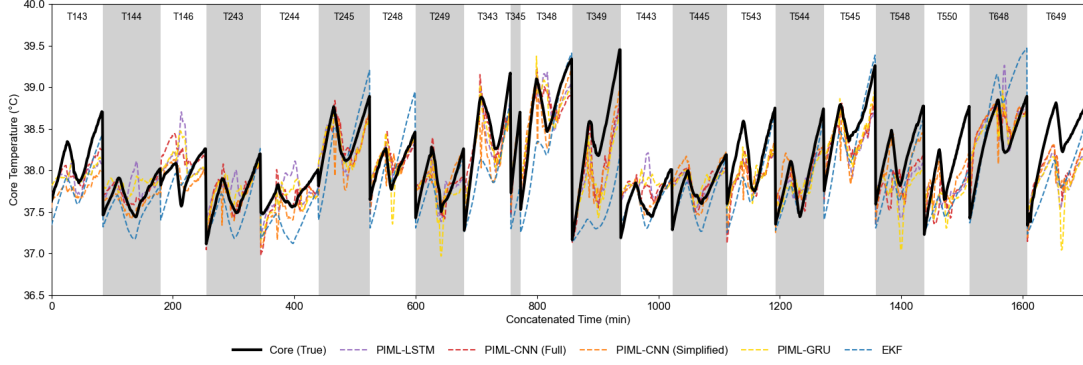
Noise Robustness Evaluation: Noise robustness evaluation examines how well a model maintains its estimation accuracy when the input signal is affected by measurement noise. To assess robustness against heart rate noise—which is common in wearable sensing—artificial Gaussian noise with zero mean and standard deviations $\sigma \in \{5, 10, 15, 20, 30, 50\}$ bpm was added to the HR signal in the test set. All models were re-evaluated under each noise condition using the same reference ground truth.

High-Risk CBT Identification: To further assess each model’s ability to identify high-risk CBT episodes, the receiver operating characteristic (ROC) analysis was conducted. A threshold of 38.0°C was adopted, consistent with occupational heat stress screening guidelines [181]. Binary classification performance was quantified using the area under the ROC curve (AUC), sensitivity (true positive rate), and specificity (true negative rate). These threshold-based metrics provide practical insight into each model’s suitability for early warning and real-time decision support in occupational and exercise heat stress monitoring scenarios.

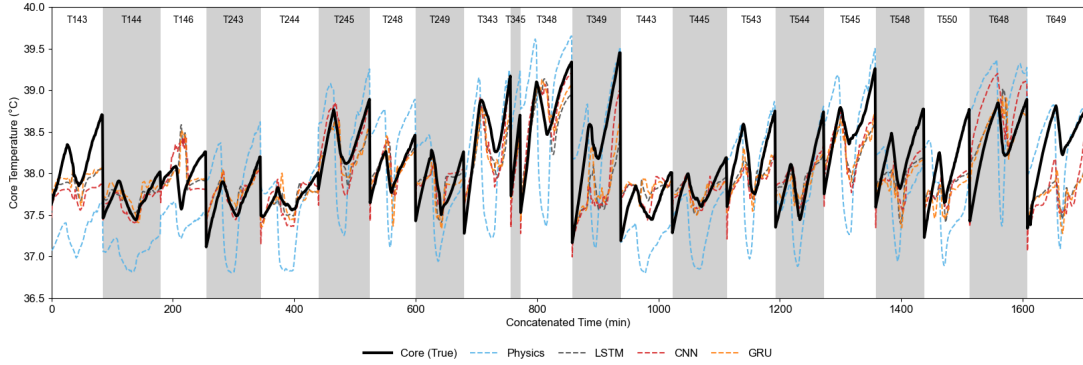
7.3 Results

7.3.1 Overall Estimation Performance

Figure 7.3 presents the time-series comparison between estimated and reference CBT across the test participants, and Table 7.3 summarizes the quantitative performance of all evaluated models. Across all metrics, the proposed physics-guided hybrid models achieve markedly higher accuracy and temporal consistency than both purely data-driven and physics-only approaches.



(a) Hybrid models and EKF



(b) Data-driven and physics-only models

Figure 7.3: Time-series comparison between estimated and reference core temperatures across the test set. (a) Hybrid models (Hybrid-CNN, Hybrid-LSTM, Hybrid-GRU) and the EKF model. (b) Purely data-driven networks (CNN-only, LSTM-only, GRU-only) and the physics-only model. Grey/white alternating background is used to visually separate different trials. Each trial is labelled using a three-digit identifier, where the first digit denotes the experimental condition and the final two digits correspond to the original participant number.

Table 7.3: Performance comparison of all evaluated models on the test set. Bold values indicate the best performance in each column.

Model	RMSE (°C)	MAE (°C)	MD (°C)	SD (°C)	R^2	DTW
Hybrid-CNN (full input)	0.265	0.202	-0.076	0.254	0.638	6.525
Hybrid-CNN (simplified input)	0.268	0.209	-0.112	0.243	0.632	6.335
CNN-only	0.322	0.243	-0.100	0.306	0.467	5.991
Hybrid-LSTM	0.310	0.239	-0.065	0.303	0.557	7.594
LSTM-only	0.323	0.240	-0.094	0.309	0.519	8.266
Hybrid-GRU	0.312	0.242	-0.080	0.301	0.551	7.925
GRU-only	0.316	0.239	-0.100	0.300	0.538	8.031
EKF model	0.386	0.300	-0.226	0.313	0.234	7.549
Physics-only model	0.604	0.520	-0.149	0.586	-0.877	13.147

The hybrid CNN models yielded the best overall performance. The full-input Hybrid-CNN achieved the lowest RMSE (0.265 °C) and MAE (0.202 °C), as

well as the highest coefficient of determination ($R^2 = 0.64$). Its simplified-input counterpart—using only heart rate and skin temperature—produced nearly identical results (RMSE = 0.268 °C, $R^2 = 0.63$), demonstrating that ambient temperature and humidity sensors can be omitted without substantial accuracy loss.

All hybrid models achieved better performance than their corresponding purely data-driven versions (LSTM-only, GRU-only, and CNN-only), indicating the effectiveness of physics-guided residual learning in improving both prediction accuracy and temporal coherence.

The EKF model achieved moderate performance (RMSE = 0.386 °C, $R^2 = 0.24$), maintaining reasonable temporal smoothness but showing larger bias (MD = −0.23 °C) and lower correlation with the reference temperature. The purely physics-based formulation exhibited the largest errors (RMSE = 0.60 °C, negative R^2), underscoring the limitations of static parameter assumptions and the necessity of adaptive correction.

Figure 7.4 compares representative models from four methodological categories: the physics-only formulation, the Extended Kalman Filter (EKF), a purely data-driven CNN, and the proposed Hybrid-CNN (full input) model. The Hybrid-CNN demonstrates the most compact and symmetric residuals centered near zero, indicating minimal bias and high consistency across participants. In contrast, the EKF model exhibits a broader residual spread with a clear tendency toward underestimation, while the purely data-driven CNN shows a moderately wider spread and slightly higher variance. The physics-only model performs the poorest, with large dispersion and a systematic bias. These residual characteristics are consistent with the quantitative results in Table 7.3, confirming that the hybrid approach effectively reduces both random error and systematic bias, thereby providing more stable and physiologically coherent temperature estimates.

Bland–Altman analysis was also performed on these four representative models. Figure 7.5 presents the Bland–Altman plots comparing the predicted and reference core temperatures across all test participants.

The Hybrid-CNN (full input) exhibits the closest agreement with reference measurements, showing the narrowest limits of agreement (LoA) and the smallest mean bias. The purely data-driven CNN shows slightly wider LoA, indicating a

mild loss of consistency compared with its hybrid counterpart. The EKF model demonstrates broader spread and a clear tendency toward underestimation, while the physics-only model performs the poorest, with the largest bias and variance. Overall, the hybrid approach effectively reduces both random and systematic errors, achieving high agreement with the ground-truth core temperature across diverse subjects and environments.

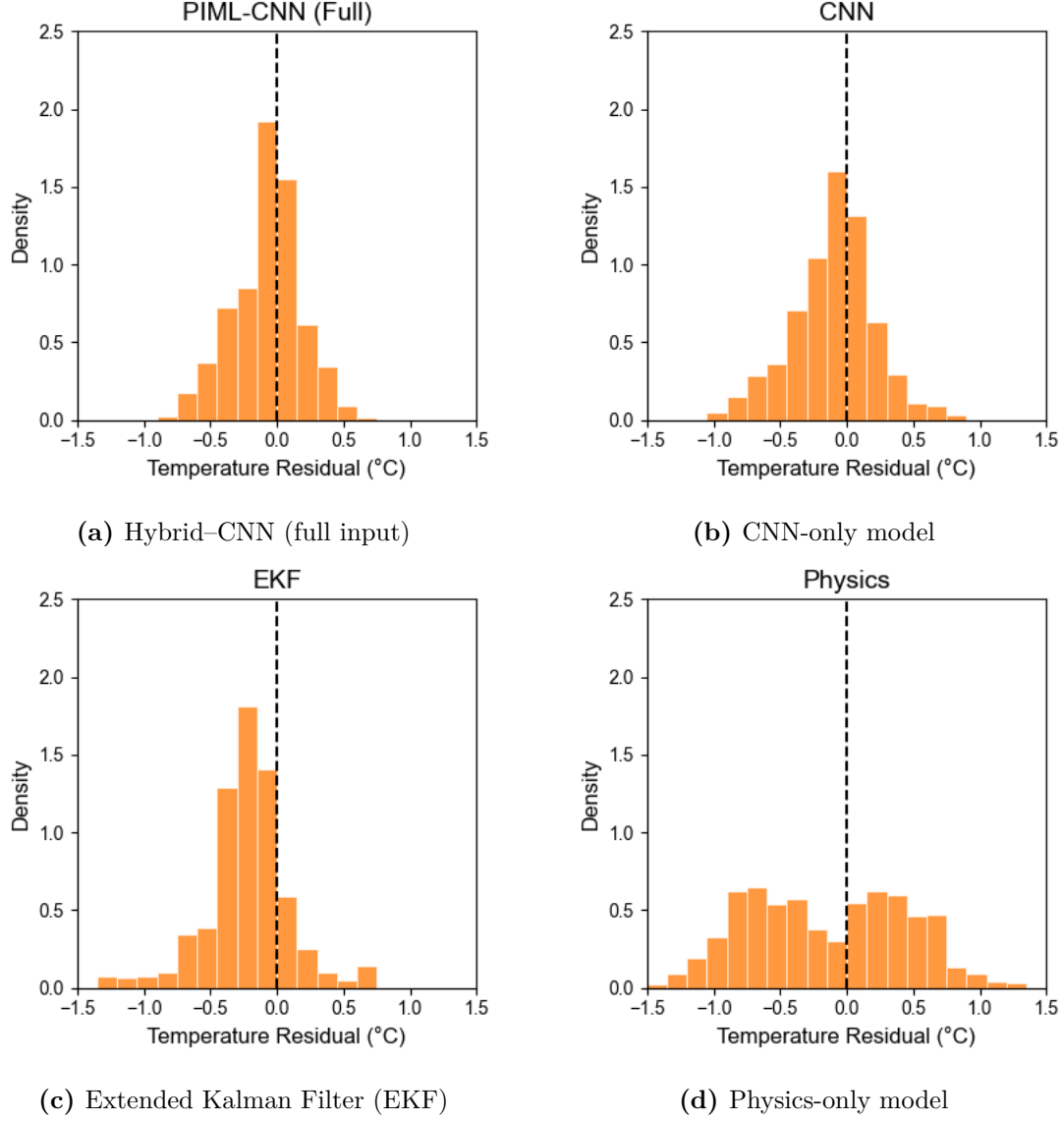


Figure 7.4: Residual distributions ($T_{\text{pred}} - T_{\text{true}}$) for representative models. The Hybrid-CNN (full input) shows the most compact and symmetric residuals centered near zero, The EKF model shows wider spread and underestimation, The purely data-driven CNN displays slightly wider spread. The physics-only model performs the poorest, with large dispersion and evident bias.

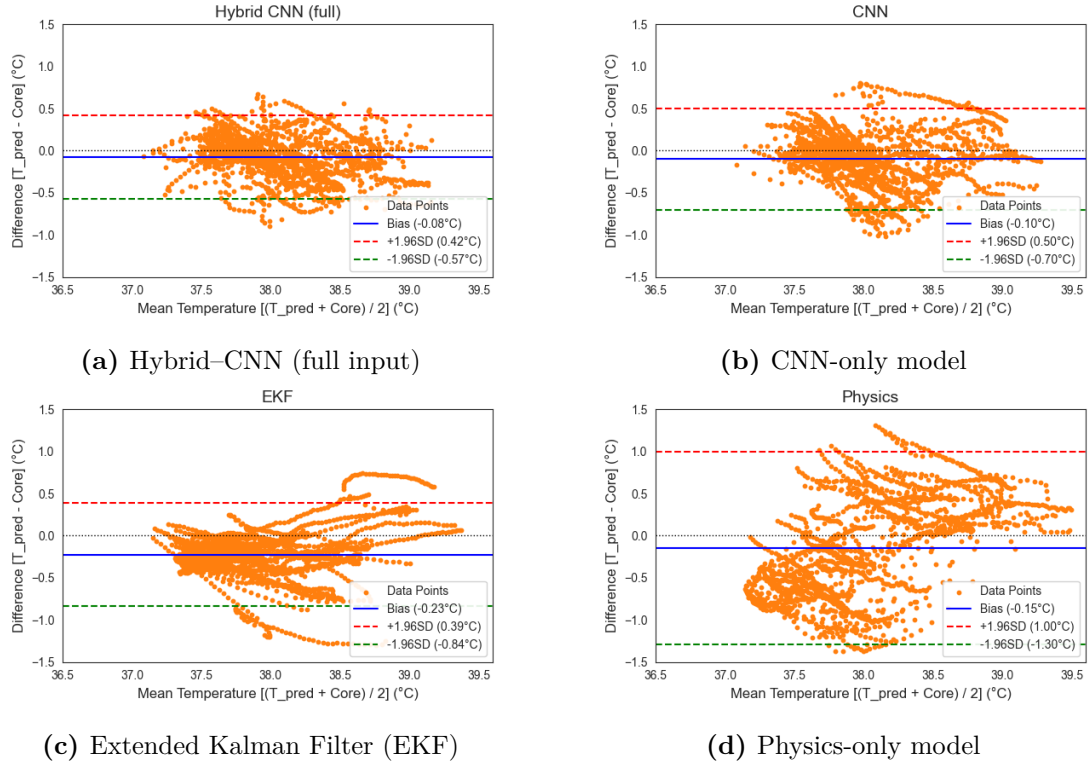


Figure 7.5: Bland–Altman plots comparing predicted and reference core temperatures for representative models. The Hybrid–CNN (full input) achieves the narrowest limits of agreement and minimal bias. The purely data-driven CNN shows slightly wider limits and higher variance. The EKF model exhibits clear underestimation, while the physics-only model demonstrates the largest bias and poorest agreement overall.

To evaluate the practical capability of each model for detecting high-risk thermal conditions, a threshold-based analysis was conducted using a cutoff of 38.0 °C. Predicted core temperatures were binarized according to this threshold, and standard classification metrics—including the area under the ROC curve (AUC), sensitivity, specificity, and overall accuracy—were computed on the test set.

Table 7.4: Threshold-based detection performance ($T_c \geq 38.0^\circ\text{C}$) on the test set, sorted by sensitivity.

Model	AUC	Sensitivity	Specificity	Accuracy
Hybrid–CNN (full input)	0.903	0.765	0.886	0.821
Hybrid–CNN (simplified)	0.915	0.709	0.888	0.792
Hybrid–GRU	0.854	0.677	0.900	0.781
Hybrid–LSTM	0.859	0.672	0.880	0.769
LSTM-only	0.869	0.627	0.924	0.765
CNN-only	0.839	0.613	0.917	0.754
GRU-only	0.865	0.604	0.933	0.757
EKF model	0.918	0.589	0.990	0.776
Physics-only model	0.790	0.604	0.745	0.669

As summarized in Table 7.4, the proposed Hybrid-CNN (full input) achieved the highest sensitivity (0.77) and overall accuracy (0.82), while maintaining high specificity (0.89). Its simplified version, using only heart rate and skin temperature as inputs, retained comparable performance. Among the baseline models, the purely data-driven CNN and EKF showed stronger bias toward specificity and lower sensitivity, resulting in missed detections under elevated temperature conditions.

To assess robustness under realistic wearable sensing conditions, additive Gaussian noise was injected into the heart-rate (HR) signal during testing. All models were evaluated on the same test set, to which different levels of noise were added, with standard deviations of $\sigma = 0, 5, 10, 15, 20, 30$, and 50 bpm. All other input signals and model parameters were kept identical across conditions, ensuring a fair and controlled comparison of noise robustness. This analysis simulates practical scenarios in which HR measurements are affected by motion artifacts, poor skin contact, or transient signal dropouts.

Table 7.5 summarizes the RMSE performance of all models under different HR noise levels. As illustrated in Figure 7.6, the proposed Hybrid-CNNs achieve the highest accuracy for low-to-moderate noise levels ($\sigma \leq 20$ bpm). When the HR signal degradation becomes severe ($\sigma = 50$ bpm), the Hybrid-LSTM surpasses all other methods (RMSE = 0.3532 °C), demonstrating higher robustness to signal corruption.

Table 7.5: RMSE (°C) under injected HR noise of varying standard deviation σ . All models were evaluated on the same test set with identical noise conditions.

Model	0	5	10	15	20	30	50
Hybrid-CNN (full input)	0.2652	0.2680	0.2790	0.2872	0.3053	0.3357	0.4072
Hybrid-CNN (simplified input)	0.2675	0.2614	0.2557	0.2819	0.3140	0.4006	0.5576
Hybrid-LSTM	0.3097	0.3100	0.3104	0.3172	0.3266	0.3351	0.3532
Hybrid-GRU	0.3118	0.3123	0.3152	0.3169	0.3207	0.3517	0.3794
CNN-only	0.3220	0.3319	0.3556	0.3721	0.3805	0.4322	0.4880
LSTM-only	0.3228	0.3244	0.3222	0.3348	0.3362	0.3487	0.3776
GRU-only	0.3163	0.3185	0.3213	0.3423	0.3535	0.3961	0.4849
EKF model	0.3860	0.3859	0.3871	0.3810	0.3871	0.3945	0.3863
Physics-only model	0.6042	0.6044	0.6057	0.6047	0.6112	0.6104	0.6332

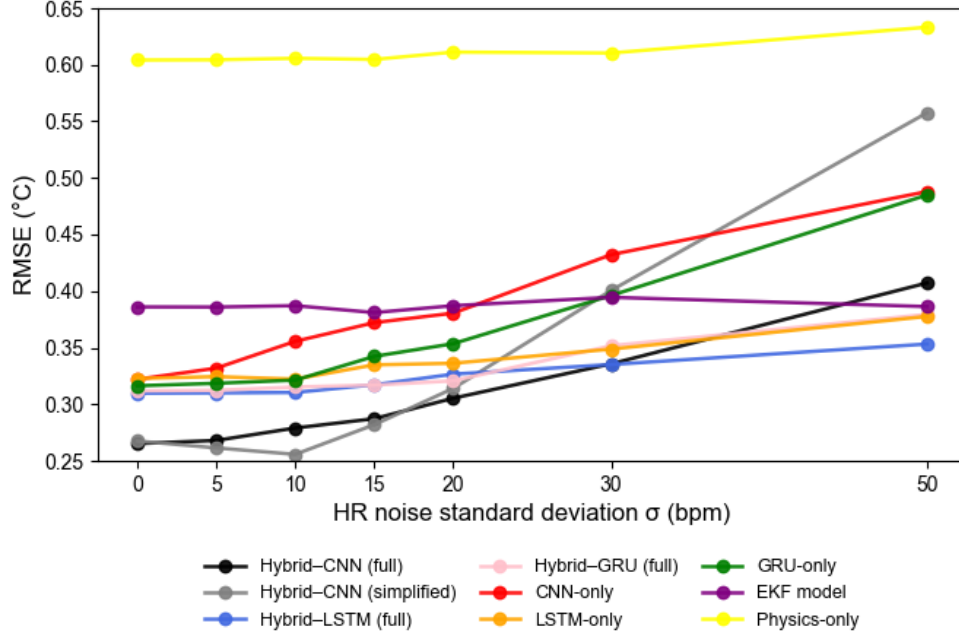


Figure 7.6: Noise-robustness analysis under injected Gaussian noise on HR input. Hybrid-CNNs achieve the lowest RMSE for low-to-moderate noise levels ($\sigma \leq 20$ bpm), while Hybrid-LSTM achieves the best performance under extreme noise ($\sigma = 50$ bpm).

7.4 Discussion

This study presents a physics-guided hybrid modeling framework that integrates a two-node thermoregulation model with deep neural residual learners for non-invasive estimation of core body temperature (CBT) from wearable sensor data. Comprehensive comparisons across multiple model architectures and baseline methods—including physics-only formulations, Extended Kalman Filters (EKF), and standalone neural networks—demonstrate that combining first-principles physiological modeling with data-driven residual correction substantially improves estimation accuracy and temporal stability. By embedding interpretable thermal dynamics within a flexible learning framework, the proposed hybrid approach achieves consistent performance across diverse participants and environmental conditions while maintaining robustness to measurement noise and physiological variability.

Comparisons with baseline models further underline the strength of the proposed hybrid framework over both physics-only and purely data-driven approaches. Across all architectures, the hybrid models consistently achieved better performance than

their standalone neural counterparts, demonstrating that coupling physiological priors with data-driven residual learning yields more accurate and stable temperature estimates. Among these, the Hybrid-CNN achieved the lowest overall estimation error and maintained the most compact residual distribution, confirming its effectiveness in capturing localized temporal-spatial patterns while preserving physiological consistency. While not the top performer under nominal conditions, the Hybrid-LSTM exhibited remarkable resilience under severe heart rate noise perturbations, surpassing the EKF when the signal degradation became extreme ($\sigma = 50$ bpm). This finding suggests that recurrent residual structures can still provide strong temporal filtering capability when faced with high-frequency noise or corrupted input streams. In contrast, the EKF exhibited strong robustness due to its probabilistic update mechanism but remained fundamentally constrained by its linearized dynamics and limited adaptability to nonlinear physiological variations. Overall, these comparisons indicate that embedding interpretable thermoregulatory priors within a deep learning framework provides a clear and systematic advantage in both estimation accuracy and robustness over conventional filtering and standalone neural models.

Beyond continuous regression metrics, a threshold-based evaluation was performed to assess each model’s ability to identify high-risk thermal states ($T_c \geq 38.0^\circ\text{C}$), an occupationally recognized cutoff for heat stress [181]. Although the EKF achieved the highest AUC, it favored conservative predictions with high specificity but reduced sensitivity, while purely neural models similarly showed limited sensitivity in detecting high-risk conditions. In contrast, the proposed Hybrid-CNN achieved the most balanced detection performance, combining strong sensitivity with stable overall accuracy, and its simplified variant retained comparable discriminative capability using only heart rate and skin temperature inputs. Such balance between physiological interpretability and predictive reliability is particularly valuable for wearable or field applications, where minimizing false negatives (i.e., undetected heat stress) is far more critical than occasional false alarms. Overall, integrating physics-based thermoregulation with data-driven residual learning enhances not only continuous estimation but also the robustness of threshold-based decision-making in safety-critical contexts.

The improved performance of the hybrid framework can be attributed to the complementary strengths of its two components. The two-node thermoregulation model provides a physiologically grounded foundation that captures the essential heat transfer dynamics between the body core and skin under varying metabolic and environmental conditions. This structured prior constrains predictions within a physically plausible range and offers a reliable initialization for subsequent data-driven refinement. In parallel, the neural residual learner focuses on modeling unaccounted variability—such as inter-individual differences in thermoregulatory efficiency, transient physiological responses, and sensor-induced fluctuations—that cannot be fully represented by simplified biophysical equations. It is also important to emphasize that in the proposed hybrid framework, the physics-based model determines the overall structure and trajectory of CBT dynamics, while the neural network is used solely to model residual discrepancies between the physics-derived prediction and the true CBT, rather than directly generating temperature estimates. This separation preserves the physiological interpretability and thermodynamic consistency of the hybrid system. Unlike purely data-driven approaches, the model maintains a clear division of responsibility: the physics-based layer encodes biophysical principles, whereas the neural component enhances adaptability to unmodeled and individualized effects. This cooperative mechanism underpins the framework’s strong generalization across subjects and environmental settings, as well as its resilience against corrupted or noisy inputs.

Several modeling considerations further clarify the observed behaviours and implementation choices. First, the metabolic heat production rate (M) is estimated directly from heart rate (HR) using an established empirical relationship, allowing the model to infer internal heat generation continuously from a single physiological signal. Second, instead of estimating skin temperature dynamically through thermodynamic balance, as in traditional two-node models, the measured head skin temperature from wearable sensors is directly integrated. This modification improves model responsiveness during transient events and aligns with the practical realities of wearable deployment, where real-time peripheral temperature data are readily available. The assumption of zero external mechanical work ($W = 0$) is also

justified, as the majority (typically over 75–80%) of metabolic energy produced during physical activity is dissipated as heat rather than converted into external work [263]. Although this simplification may cause slight overestimation of core temperature during intense exercise, it is physiologically reasonable and essential for practical implementation in real-world wearable systems. In future implementations, this limitation could be mitigated by incorporating auxiliary sensor modalities such as accelerometer or gyroscope data to infer external mechanical work in real time, thereby enabling dynamic adjustment of the effective metabolic heat load within the hybrid thermoregulation framework.

The model also assumes a standardized body configuration representing a 30-year-old individual with a body mass of 70 kg and 1.8 m² body surface area, consistent with the demographic profile of typical heat-stress-exposed adults such as athletes, military personnel, and outdoor workers. While this parameterization inevitably abstracts inter-individual variability, it reflects a deliberate design choice that balances physiological realism with generalizability. Human thermoregulation is inherently complex and influenced by numerous interacting factors; therefore, highly detailed or individualized parameterization often reduces robustness and increases overfitting risk when applied across diverse populations. In contrast, a simplified, population-representative configuration provides a stable foundation that generalizes well to individuals performing comparable physical tasks in similar environments. However, this modeling approach does not explicitly account for sex-based physiological differences—such as variations in body mass and height—which may limit its applicability to female populations and warrant further investigation in future work.

To assess robustness, heart rate (HR) noise was specifically considered, as it represents a realistic challenge in wearable applications. In both Dataset 1 and Dataset 2, HR was obtained from wearable chest electrocardiogram (ECG) recordings, which generally provide more reliable measurements during physical activity than commercial wrist-based photoplethysmography (PPG) devices. Nevertheless, in practical wearable deployments, HR signals—particularly those obtained from PPG-based sensors—are susceptible to motion artifacts, poor contact, and physiological variability [269, 270], while skin and ambient temperatures are typically recorded

with relatively high fidelity. In fact, commercial wearable devices may exhibit mean absolute errors (MAE) of 10–20 beats per minute (bpm) under physical activity conditions [271]. By injecting synthetic Gaussian noise of varying magnitudes into the HR input, we emulated the signal perturbations commonly encountered in real-world deployments. The results showed that the Hybrid-CNN maintained the highest accuracy across low to moderate noise levels ($\sigma \leq 20$ bpm), while the Hybrid-LSTM became more resilient as noise increased, reflecting the advantage of recurrent temporal modeling under degraded input conditions. These findings underscore the framework’s potential for practical wearable use, where HR noise remains one of the dominant sources of uncertainty.

Finally, the comparable performance between the full-input and simplified Hybrid-CNN models under nominal conditions provides further insight into model structure. Because skin temperature was directly measured from wearables, the ambient air temperature (T_a) and relative humidity (RH) contributed limited additional information to the skin–environment heat exchange in the two-node model. However, the neural residual learner receives both physics-based predictions and these environmental variables as inputs, allowing it to implicitly capture the coupling between skin and environment. This explains why, under high-noise conditions, the full-input Hybrid-CNN achieved better performance than the simplified variant: when HR becomes unreliable, the model increasingly relies on skin temperature and environmental inputs to maintain accurate thermal state estimation.

Notably, this study relies solely on head skin temperature as a proxy for peripheral thermal status. While conventional multi-site methods (e.g., seven-point weighted averages [272]) can better approximate mean skin temperature, such configurations are impractical for continuous, wearable monitoring due to comfort, cost, and power constraints. Using a single head-mounted sensor therefore represents a pragmatic compromise between physiological representativeness and deployability. Although localized gradients may cause small discrepancies between head and whole-body skin temperature, prior studies indicate that forehead and trunk regions exhibit the highest stability and strongest correlation with core thermal status [273, 274]. Future

work may extend this framework to include sparse multi-site fusion or adaptive correction mechanisms to further improve peripheral temperature representation.

While ambient variables can improve CBT estimation accuracy, obtaining reliable environmental measurements from wearable devices remains challenging. Close-proximity sensors may be affected by body heat, sweat, and evaporative cooling—especially during physical exertion—leading to measurement bias. In this study, ambient temperature and humidity were recorded by external instruments located within a controlled climate chamber, which ensured accurate environmental readings. In practical applications, such biases can be mitigated by physically distancing the sensor from the body, such as placement near the collar, helmet, or inside a ventilated enclosure. Additional solutions include passive thermal shielding, airflow-based isolation, or real-time calibration and filtering techniques. In addition, in many large-scale deployments—such as factories, athletic competitions, or occupational safety monitoring—environmental conditions are relatively homogeneous within a given area. In such contexts, centralized environmental monitoring will be sufficient, eliminating the need for individual ambient sensors and simplifying system integration.

From a computational standpoint, the proposed hybrid framework is well suited for real-time wearable deployment. As summarized in Table 7.2, all neural residual learners are lightweight, with parameter counts below 3.5×10^4 and model sizes ranging from approximately 120 to 140 KB. The total computational cost per inference step remains modest, between 250K and 700K FLOPs, corresponding to a median wall-clock latency of about 50 ms when executed on a consumer laptop equipped with an AMD Ryzen 7 processor. Because the physics-based two-node model is purely algebraic and operates in an open-loop configuration, it contributes negligibly to the overall computational burden—its execution time is well below the neural forward-pass latency. Consequently, the real-time performance of the hybrid system is effectively governed by the residual learner, with no appreciable delay introduced by the physical component.

These characteristics confirm that the proposed framework achieves a favorable balance between accuracy and computational efficiency. While classical filtering methods such as the Extended Kalman Filter (EKF) provide the lowest latency, their

accuracy is substantially lower. In contrast, the hybrid models deliver state-of-the-art estimation performance with comparable complexity than typical deep learning architectures. With total memory requirements under 150 KB and per-step inference times around 50 ms. The framework comfortably meets the computational constraints of modern microcontrollers and embedded processors, confirming its practical viability for continuous, on-device core temperature monitoring in wearable systems.

It is also worth noting that the size and diversity of the training dataset in this study remain limited. The current evaluation primarily involves healthy individuals under controlled experimental conditions. To improve generalizability, future work should extend data collection to encompass broader populations—including varying age groups, fitness levels, and thermoregulatory profiles—and more diverse environmental exposures. Such expansion would allow for more rigorous training and validation of the residual compensation module, and may support transfer learning or domain adaptation strategies in practical deployments.

Regarding dynamic responsiveness, the behaviour of the hybrid model is primarily governed by the underlying two-node thermoregulation module. Although this physics-based structure ensures physiological consistency, its response to abrupt changes in environment or activity is inherently influenced by the body’s thermal inertia and the limited rate of heat exchange. Additionally, since the model relies on inputs such as skin temperature and heart rate—which reflect metabolic and thermal dynamics—there is a natural delay between the onset of physiological changes and their manifestation in sensor measurements. However, given the current used sampling rate of one minute, which is relatively long, the model is sufficiently responsive to capture the timescale of most real-world transient events. This sampling interval aligns well with the slower dynamics of CBT regulation and avoids excessive noise amplification. Nonetheless, future implementations could explore forward-looking strategies such as multi-step simulation based on estimated activity trends or environmental forecasts, thereby enhancing responsiveness during rapid transitions while preserving model interpretability.

Although Physics-Informed Neural Networks (PINNs) represent a popular paradigm within the physics-informed machine learning (PIML) family, this study intentionally

adopts a residual-based formulation instead of a PINN-style framework. The primary reason lies in the nature of the underlying two-node thermoregulation model, which serves as a simplified surrogate with fixed physiological parameters. This simplification introduces structural bias, as it cannot fully capture inter-individual variability, clothing effects, or transient physiological changes under exercise. Embedding such equations directly into a neural loss function, as in PINNs, would reinforce these modeling inaccuracies during training and constrain the potential for accuracy improvement. Furthermore, the limited size and diversity of the available dataset make PINN optimization particularly challenging. Training a PINN requires balancing physics and data losses under high-fidelity supervision or multiple boundary conditions, which is impractical in this small-data regime and may lead to over-regularization toward the physics term. In contrast, the residual-modeling PIML formulation used here decouples the roles of the physical and neural components: the two-node model provides a transparent baseline trajectory, and the neural learner focuses solely on correcting residual discrepancies. This design mitigates model bias, enhances data efficiency, and maintains computational feasibility for on-device execution.

Looking ahead, future studies may explore more interpretable neural architectures to enhance the transparency and adaptability of the residual learner. Emerging frameworks such as Kolmogorov–Arnold Networks (KANs) and their extensions—including fractional, rational, and wavelet-based variants—offer analytic basis functions that could serve as alternative representations within the residual-learning module or as interpretable components in hybrid PINN–PIML formulations [275–278]. Additionally, incorporating automated neural architecture search (NAS) could help identify optimal network topologies for balancing interpretability, accuracy, and computational efficiency [279].

7.5 Conclusion

This study presented a physics-guided hybrid framework that integrates a two-node thermoregulation model with neural residual learners for estimating core body temperature (CBT) from wearable sensor data. The proposed approach

combines physiological interpretability with data-driven adaptability, achieving higher accuracy and robustness than physics-only, filtering-based, and purely neural models. Among the variants, the Hybrid-CNN demonstrated the best overall accuracy under nominal and low-noise conditions, whereas the Hybrid-LSTM showed higher resilience under severe heart rate noise. The framework remains computationally lightweight, supporting real-time inference on wearable devices. Future extensions will incorporate additional sensor modalities, such as accelerometer-derived workload estimation, to dynamically account for external mechanical work and enhance physiological fidelity. Overall, this work establishes a practical and interpretable foundation for real-time thermal state monitoring in wearable health systems.

8

Conclusions and Future Work

8.1 Conclusions

This thesis investigated non-invasive core body temperature (CBT) estimation using wearable physiological signals. By developing and evaluating a series of estimation frameworks using probabilistic, adaptive, and physics-guided approaches, the reliability of non-invasive CBT estimation was improved.

Chapter 2 presented a systematic review of non-contact infrared thermometers (NCITs) and infrared thermography (IRT) for human body temperature measurement, summarizing their use across a wide range of experimental and clinical settings. The review demonstrated that, although non-contact techniques show promising potential for screening applications, their accuracy remains limited when compared with contact reference standards. Current non-contact techniques are insufficient for reliable monitoring of internal thermal strain during physical activity. Therefore, conventional contact-based methods are still necessary in these scenarios.

Chapter 3 presented a systematic review of data-driven core body temperature estimation methods using non-invasive signals during physical activities. It summarized four commonly reported method categories, including statistical models, KF/EKF filtering approaches, machine learning methods, and hybrid frameworks. In addition to the literature review, this chapter conducted a benchmark evaluation of selected methods using two independent datasets. The benchmark validation results varied across studies, and in several cases differed from the performance reported in the original publications. These results showed that modeling choices, input requirements, and validation settings play an important role in estimation

performance. This chapter also provided a methodological basis for the developments presented in the subsequent chapters.

Chapter 4 extended classical EKF-based CBT estimation by introducing a multi-model Kalman filtering framework with variance-based fusion. Multiple EKFs were run in parallel, with each filter representing a distinct physiological state. Each EKF produced a posterior state estimate and the corresponding estimation variance, and the final core temperature estimate was obtained by fusing the multiple EKF estimates using weights based on their posterior estimation variances. This approach showed that incorporating multiple physiological states and applying variance-based weighted fusion improved estimation accuracy compared with the single EKF approach. Both label-informed Physiological State-Specific Kalman Filters (PSSK) and unsupervised Trial Clustering-Based Kalman Filters (TCBK) were investigated, illustrating that the framework can be implemented under different levels of prior knowledge.

Chapter 5 addressed physiological state uncertainty by introducing a probabilistic state inference framework based on a Hidden Markov model (HMM) and particle filter (PF). Inspired by the methods in Chapter 4 that estimation performance depends on states related to activities, this chapter explicitly modeled uncertainty in thermoregulatory state transitions. The physiological states were represented as discrete variables, and their posterior probabilities were inferred within a probabilistic framework. Meanwhile, an HMM was coupled with a particle filter, enabling state uncertainty to be explicitly represented and propagated over time. Within this framework, the inferred state probabilities influenced the contribution of state-dependent filters, rather than relying on heuristic variance-based fusion. The results indicated that explicitly modeling physiological state uncertainty leads to improved estimation accuracy under dynamic exercise conditions.

While Chapters 4 and 5 improved estimation accuracy by introducing state-dependent observation models and regulating their contribution through heuristic fusion and probabilistic state inference, respectively, systematic estimation errors still existed within individual states due to sensor noise and modeling simplifications. Chapter 6 explored an alternative strategy by compensating the residual of the observation model. It introduced a residual-compensated adaptive Kalman filter

(RCAKF). A lightweight LSTM network was integrated into the observation model of a single EKF for residual compensation in real time. The results demonstrated that residual compensation can improve estimation accuracy compared with unaugmented EKF baselines, highlighting machine learning-based observation correction as an alternative modeling pathway to state-dependent approaches.

Chapter 7 explored a physics-guided hybrid modeling strategy as a parallel pathway to the EKF/PF methods in earlier chapters. Chapter 6 demonstrated that machine learning-based residual compensation can improve estimation accuracy within the filtering framework. However, the observation model was still based on empirical assumptions and was not constrained by physiological principles, which limited interpretability and could be sensitive to data quality and noise. To address this limitation, Chapter 7 integrated a first-principles two-node thermoregulation model with data-driven residual compensation. In this framework, the two-node thermoregulation model provides a physiological baseline, while the neural network compensates for systematic residuals of the thermoregulation model. By embedding physiological priors directly into the estimation process, the hybrid model constrains learning within a physically interpretable structure. The results showed that the physics-guided hybrid framework achieved the lowest RMSE among the evaluated methods, including the EKF and purely machine learning baselines. In addition, the hybrid model showed improved robustness under noisy heart rate signal conditions. These findings demonstrate the promising potential of physics-guided hybrid modeling as an effective and interpretable approach for wearable core body temperature estimation.

On Dataset 1, the best RMSE decreased progressively from 0.38 °C in Chapter 4 to 0.359 °C in Chapter 5, 0.31 °C in Chapter 6, and 0.27 °C in Chapter 7. This progression demonstrates that estimation accuracy was gradually improved through the introduction of physiological-state modelling, probabilistic state inference, adaptive residual compensation, and physics-guided learning. However, these values provide an overall indication of methodological progress rather than a strictly controlled ranking, because the methods differ in their model structures, assumptions, and evaluation details.

As discussed in Section 1.2, no universally accepted RMSE threshold currently exists for CBT estimation from wearable physiological signals. The ± 0.3 °C accuracy requirement specified for clinical thermometers relates to the technical accuracy of temperature measurement equipment compared with calibrated reference temperature sources and is not a formal acceptance criterion for algorithmic CBT estimation. Nevertheless, an RMSE of approximately 0.3 °C may be used as a pragmatic numerical reference for interpreting performance under controlled conditions. On this basis, the methods developed in Chapters 6 and 7 approached or achieved this numerical reference on Dataset 1. These findings demonstrate that the primary aim of improving the accuracy of non-invasive CBT estimation from limited wearable physiological signals was achieved within the scope of Dataset 1 and the evaluated experimental conditions.

However, the methods developed in this thesis differ in computational cost, model complexity, robustness, and physiological interpretability. Therefore, the method with the lowest RMSE may not necessarily be the most appropriate for every practical application. Selection should also account for the available computational resources, real-time requirements, sensor inputs, interpretability requirements, and acceptable estimation error.

The target accuracy for CBT estimation cannot currently be translated into a specific accuracy requirement for the HR measurement equipment. Rather than the uncertainty in the HR input, the reported CBT estimation error mainly reflects the limitations in model structure, parameter estimation, inter-individual variability, and the uncertainty of the reference temperature measurements. As these sources of error have not been isolated and the estimation models cannot be assumed to be optimal, the contribution of HR measurement error to the overall CBT RMSE cannot be quantified independently. A formal HR accuracy requirement would therefore require a dedicated sensitivity analysis using an optimal estimation model.

8.2 Discussion

This thesis investigated CBT estimation under wearable sensing constraints through a series of complementary studies. Rather than seeking a single optimal

solution, this work explored multiple modeling strategies with different input requirements to improve estimation performance across varying activity states and environmental conditions. In practice, estimation methods should be selected based on the available information and operational constraints, such as noise levels and device capabilities, rather than solely peak performance.

A fundamental challenge addressed in this work is the indirect and non-linear relationship between wearable physiological signals, particularly heart rate, and the internal thermal state. Heart rate is widely available and practical for wearable monitoring, but it serves as a proxy for metabolic rate and thermoregulatory processes and is influenced by multiple confounding factors. As a result, estimation errors can arise from simplified observation models, uncertainty in physiological state transitions, sensor noise, and variability across individuals.

The modeling strategies explored in this thesis illustrate how different forms of structure can be introduced to address a challenging estimation problem. State-dependent methods describe the system using multiple observation models within EKF or PF, and each observation is associated with a distinct physiological state. Probabilistic state inference explicitly models the uncertainty in physiological states and their transitions. Residual compensation approaches address systematic errors arising from simplified observation models, while physics-guided hybrid models incorporate thermoregulatory principles to provide physiologically grounded constraints. As a result, different modeling structures offer complementary mechanisms for managing uncertainty and error.

In this thesis, heart rate is used as the primary input signal as a practical design choice. In particular, Chapters 4, 5, and 6 rely on heart rate as the sole physiological input. Heart rate offers a practical balance between physiological relevance and wearable feasibility. It is widely available on commercial wearable devices and reflects cardiovascular and metabolic responses to thermal stress during physical activity. Although heart rate alone cannot capture all of the information of human thermoregulation, focusing on HR allowed us to develop practical estimation methods under realistic wearable constraints that better reflect everyday scenarios.

From a methodological standpoint, the multiple modeling pathways explored in this thesis are important for wearable heat strain monitoring. In real-world applications, practical constraints must be considered, including sensor availability, computational resources, robustness to noise, and physiological interpretability. Highly flexible data-driven models may achieve good performance under controlled conditions, but they can be sensitive to data quality and operating scenarios. In contrast, models with stronger physiological constraints usually offer greater robustness and interpretability, but with reduced flexibility. As a result, modeling choices should be guided by operational constraints and dominant sources of uncertainty, rather than by the pursuit of a single model with the highest reported accuracy.

From an application perspective, the modeling strategies explored in this thesis are particularly relevant to scenarios where continuous and convenient thermal strain monitoring is required, including occupational heat strain monitoring in hot environments, sports and training, and population-level heat risk assessment under extreme climate conditions. In such settings, robustness to sensing noise, physiological interpretability, and compatibility with existing wearable platforms may be as important as marginal gains in estimation accuracy.

8.2.1 Limitations

Several limitations of this work should be acknowledged. Firstly, the datasets used in this thesis are limited in size and diversity, which constrains the generalizability of the models. Secondly, the reliance on heart rate as the primary input signal reflects practical wearable considerations but limits the information available to the models. Finally, as the proposed methods rely on supervised learning, uncertainty in reference core body temperature measurements during physical activity also limits the achievable estimation accuracy.

Meanwhile, the models were trained and evaluated under controlled experimental conditions. While this enabled systematic comparison across methods, it does not fully capture the variability encountered in long-term free-living applications, such as changes in sensor placement, user behaviour, and device characteristics [280]. Future work should therefore examine the generalizability of these approaches across different devices, sampling rates, and environmental and activity conditions, and

explore adaptive and personalized strategies suitable for long-term deployment in free-living settings.

In addition, the available datasets primarily included healthy participants under controlled experimental conditions. This is consistent with the main application scenarios considered in this thesis, including occupational heat-strain monitoring, sports and training, and heat-risk assessment in generally healthy populations. Therefore, the use of healthy participants was appropriate for evaluating the proposed methods within the intended scope of the work. Nevertheless, the findings cannot be assumed to generalize directly to clinical populations. In particular, cardiac arrhythmias, such as atrial fibrillation, may produce irregular heart-rate patterns and alter the relationship between heart rate and core body temperature assumed by the estimation models [281, 282]. Further validation and, where necessary, model adaptation would therefore be required before these methods could be applied to patients with cardiovascular conditions or irregular heart rhythms.

Overall, this thesis treats non-invasive CBT estimation as a problem that can be approached using different modeling strategies, rather than a single optimal solution. Each strategy addresses uncertainty from a different perspective. Together, these findings help inform how CBT estimation methods can be selected and applied under practical wearable sensing constraints.

8.3 Future Work

Several directions remain open for future research to extend the findings of this thesis and advance wearable CBT estimation toward reliable real-world heat strain monitoring.

8.3.1 Expansion of Dataset Size and Diversity

A primary limitation of the present work is the size and diversity of available datasets. The datasets used in this thesis are limited in participant numbers, settings, and protocol diversity, which constrains the generalizability of the methods. This limitation was also emphasized in the systematic review in Chapter 3. The validation results varied across studies, and in several cases differed from the performance

reported in the original publications. Importantly, these differences should be interpreted cautiously. They do not necessarily indicate that the original methods are ineffective, but may instead reflect differences in dataset characteristics, experimental conditions, and validation designs. These observations reinforce the view that dataset quality, diversity, and representativeness can be as important as the model design. Future work should focus on collecting larger and more diverse datasets across different populations, environments, and activity conditions. This would allow for more reliable model training and a clearer assessment of how well different methods generalize to real-world use.

8.3.2 Advanced Data-Driven Models

In parallel with collecting larger datasets, future work could explore data-driven approaches that benefit from the data size, where increased data availability enables the learning of more complex relationships beyond the models considered in this thesis. A promising direction is to develop foundation models trained on large scale wearable physiological datasets, enabling the learning of general representations capturing temporal and cross-individual patterns in thermoregulatory responses. Such models could be pre-trained on heterogeneous, weakly labelled or unlabelled data, and subsequently adapted to CBT estimation tasks under limited ground-truth availability [283, 284].

Another promising direction is to make better use of large datasets to learn thermoregulatory behaviour directly from data, rather than relying solely on first-principles models. Traditional thermoregulation models are usually based on simplified assumptions and population-level parameters, which may not fully capture individual differences. In contrast, data-driven approaches can learn more flexible relationships across physiological signals and environmental conditions. In this way, machine learning can be used not only to correct model errors, but also to reveal patterns in thermoregulation that are difficult to describe using analytical models [285].

Importantly, both directions raise new challenges that align closely with the findings of this thesis. Models trained on large-scale data must be evaluated carefully with respect to distribution shift, sensor noise, inter-individual variability, and deployment constraints [286]. As dataset size increases, considerations of robustness,

interpretability, and computational efficiency become as important as estimation accuracy, particularly for wearable and edge-based applications.

8.3.3 Personalization and Online Learning

Another limitation of this work is that the estimation models were trained at the population level. This choice was mainly driven by the limited size of the available datasets and the need to ensure stable and consistent evaluation across participants. As a result, individual differences in thermoregulatory responses were captured only at the population level. With access to larger datasets containing longer recordings per individual, future work could explore different levels of personalization. At a basic level, demographic and anthropometric information, such as sex, age, body mass, or fitness-related indicators, could be included to account for systematic differences between individuals. While this may help reduce population-level bias, such variables alone are unlikely to fully represent individual thermoregulatory behaviour. More effective personalization would require adapting models to each individual based on their own historical data, allowing parameters or model components to be gradually refined for a given user.

Beyond differences between individuals, thermoregulatory responses within the same individual can also change over time. Factors such as heat acclimation, changes in fitness or health status, and daily physiological rhythms may all influence the relationship between wearable signals and core body temperature. In this thesis, all models were trained offline and remained fixed during use, which means their performance may degrade over time as these underlying conditions change. Future work could therefore investigate online or continual learning strategies that allow models to adapt during deployment [287, 288]. Such approaches could update model parameters incrementally using newly observed data, either by refining residual compensation modules (Chapter 6) or by adjusting physiological parameters in physics-guided models (Chapter 7). Allowing estimation models to evolve with the user may help maintain accuracy over long-term use and support more practical, personalized wearable monitoring.

8.3.4 Generalization to Cold Environments

Another open question is whether the estimation strategies developed in this thesis can be applied beyond hot and heat-stress conditions. The models were primarily designed and evaluated for exercise in warm-to-hot environments, where thermoregulation is mainly driven by increased metabolic heat production and heat dissipation. Their behaviour under cold exposure was not examined in this work and cannot be assumed to be similar. In cold environments, the dominant physiological responses differ from those in hot environments, including peripheral vasoconstriction and shivering for heat production [289, 290]. These mechanisms change the relationship between wearable physiological signals and internal thermal state. As a result, models trained under heat-stress conditions are unlikely to transfer reliably to cold environments without modification. Therefore, future work should evaluate the model performance in cold conditions and, if needed, develop alternative formulations or adaptations for cold environment.

8.3.5 Multimodal Sensing Optimization

Another possible direction is to explore whether additional sensing information can be incorporated without significantly increasing device complexity and cost. This thesis primarily focuses on heart rate and limited skin temperature measurements to reflect the practical constraints of wearable sensing. Future work could examine the added value of other commonly available signals, such as accelerometer data or multiple skin temperature sites, and determine whether simple multimodal combinations can lead to meaningful performance improvements [291, 292]. Improved accuracy should be balanced against the added sensing burden, so that the approach remains suitable for everyday wearable applications.

8.3.6 Integrated Operational Early-Warning Systems

Beyond CBT estimation itself, an important direction for future work is to move toward practical early-warning systems for exertional heat illness (EHI) and exertional heat stroke (EHS). While CBT is a key indicator of internal thermal strain, real-world risk assessment also depends strongly on environmental and activity

context. Combining estimated CBT with environmental heat metrics, such as wet-bulb globe temperature (WBGT), as well as ambient conditions including air temperature, humidity, solar radiation, and wind, could provide more informative indicators of heat risk in large-scale events and occupational settings [22, 293]. Such integration would help translate wearable-based CBT estimation into practical decision-support tools, for example to support event planning, work–rest scheduling, and threshold-based safety protocols [211].

In addition, although CBT is a key variable for monitoring heat strain, the risk of EHI in practice is influenced by multiple physiological indicators. Future work could therefore explore early-warning models that combine CBT estimation with other wearable-derived indicators, such as cardiovascular strain metrics from heart-rate dynamics (e.g., heart rate variability), as well as features related to the gait and movement. By combining thermal, cardiovascular, and mechanical information, these multimodal approaches may provide a more complete picture of physiological stress and improve early-warning performance, particularly in real-world field settings with diverse populations and uncontrolled conditions.

8.3.7 Extension to Other Applications

Although the methods developed in this thesis focus on CBT estimation, the underlying computational strategies may be applicable to a broader range of physiological and engineering state-estimation problems. In physiological monitoring, potential applications include the estimation of fatigue and recovery status, metabolic demand, and thermal comfort from wearable signals. The methods may also be extended to industrial systems in which important internal states cannot be measured directly and must instead be inferred from available sensor measurements. For example, EKFs and PFs have been widely applied to battery state of charge, state of health, and remaining useful life estimation [294, 295]. The state-dependent and probabilistic approaches developed in Chapters 4 and 5 may be useful for systems operating under different regimes, while the adaptive residual-compensation framework introduced in Chapter 6 may help address the model mismatch. Similarly, the physics-guided approach developed in Chapter 7 could be applied when a mechanistic model cannot fully represent the variability of the system. Such extensions would require specific

process and observation models, appropriate sensor data, and independent validation. Therefore, the current CBT estimation models cannot be transferred directly, but the general principles of state-dependent modelling, probabilistic inference, adaptive correction, and physics-guided learning may support a wider range of physiological and industrial monitoring applications.

Appendices



Data and Reproducibility

This appendix provides information on data availability and a consolidated guide to the algorithmic descriptions supporting the methods developed in Chapters 4–7.

A.1 Dataset Availability

Dataset 1 is publicly available from the Loughborough University Research Repository at:

<https://doi.org/10.17028/rd.lboro.26076577.v1>

Dataset 2 is not publicly available because of data-use restrictions. Access may be requested from the corresponding author of the original publication [152], subject to the applicable data-sharing conditions.

A.2 Chapter 2–3: Supplementary Study Characteristics

The systematic review presented in Chapter 2 was based on detailed data extraction from all included studies. The extracted information is provided in Supplementary Table S1 of the corresponding published article.

The supplementary table includes the measurement scenario, infrared device brand, measurement location, measurement distance, participant characteristics (group size, age, and sex distribution), ambient temperature, reference temperature distribution, reference temperature measurement site, reference measurement device, and the reported performance metrics (e.g., mean difference, standard deviation).

The complete Supplementary Table S1 is available online as supplementary material at:

<https://www.mdpi.com/article/10.3390/s23177439/s1>

The additional study characteristics extracted in the systematic review presented in Chapter 3 are provided in the supplementary table of the corresponding published article.

The supplementary table includes the environmental temperature and humidity, measurement device details (e.g. the manufacturer and model), sampling rate, experiment setting, participant characteristics (group size, age, and sex distribution).

The complete supplementary table of Chapter 3 is available online as supplementary material at:

<https://doi.org/10.1016/j.buildenv.2026.114591>

A.3 Chapter 4: PSSK and TCBK

The source code for the PSSK and TCBK methods developed in Chapter 4 is available in the following public GitHub repository:

<https://github.com/Oxford-NIL/PSSK-TCBK>

The repository provides the model implementations, data-processing and evaluation scripts in Chapter 4. Access to the original datasets remains subject to the availability and data-sharing conditions described in Section A.1.

A.4 Chapters 5–7: Algorithmic Descriptions

Detailed mathematical formulations and pseudocode for the method developed in Chapter 5 are provided in Algorithm 4.

The corresponding formulations and pseudocode for the methods presented in Chapters 6 and 7 are already included in their respective chapters and are referenced here: the RCAF framework is described in Algorithm 2, and the physics-guided hybrid framework is presented in Algorithm 3.

Algorithm 4 HMM–PF with IGA Resampling for Real-Time Core Body Temperature Estimation

Require: HR observations $\{y_t\}_{t=1}^T$, where y_t is the measured heart rate at time t

Require: Number of particles N ; process-model parameters a_0 , a_1 , and Q

Require: State-dependent observation models $\{h_s(\cdot), R_s\}_{s \in \{0,1\}}$, where $s = 0$ denotes recovery and $s = 1$ denotes exercise

Require: HMM transition matrix $\mathbf{A} = [A_{ij}]$, where $A_{ij} = P(s_t = j \mid s_{t-1} = i)$

Require: Resampling threshold $N_{\text{th}} = N/2$

Ensure: CBT estimates $\{\hat{x}_t\}_{t=1}^T$ and posterior state probabilities $\{\mathbf{p}_t\}_{t=1}^T$

Notation:

$x_t^{(n)}$ and $w_t^{(n)}$ denote the state and normalized weight of particle n at time t

$\mathbf{p}_t = [p_t(0), p_t(1)]$ denotes the posterior HMM state belief

$\mathcal{N}(z; \mu, \sigma^2)$ denotes the Gaussian probability density evaluated at z

Step 1: Initialization

for $n = 1, \dots, N$ **do**

Draw $x_0^{(n)} \sim \mathcal{N}(37.0, 0.2^2)$

Set $w_0^{(n)} \leftarrow 1/N$

end for

Initialize the HMM state belief as $\mathbf{p}_0 \leftarrow [1, 0]$

Step 2: Recursive estimation

for $t = 1, \dots, T$ **do**

1. Particle prediction

for $n = 1, \dots, N$ **do**

Draw process noise $q_t^{(n)} \sim \mathcal{N}(0, Q)$

Propagate the particle:

$$x_t^{(n)} \leftarrow a_1 x_{t-1}^{(n)} + a_0 + q_t^{(n)}$$

end for

2. HMM state prediction

Predict the physiological-state belief:

$$\mathbf{p}_{t|t-1} \leftarrow \mathbf{p}_{t-1} \mathbf{A}$$

3. State-conditioned likelihood calculation

for $s \in \{0, 1\}$ **do**

Compute the marginal observation likelihood:

$$L_t(s) \leftarrow \sum_{n=1}^N w_{t-1}^{(n)} \mathcal{N}(y_t; h_s(x_{t-1}^{(n)}), R_s)$$

end for

4. HMM posterior update

for $s \in \{0, 1\}$ **do**

Compute the unnormalized state probability:

$$\tilde{p}_t(s) \leftarrow p_{t|t-1}(s) L_t(s)$$

end for

for $s \in \{0, 1\}$ **do**

Normalize the state probability:

$$p_t(s) \leftarrow \frac{\tilde{p}_t(s)}{\sum_{r \in \{0,1\}} \tilde{p}_t(r)}$$

end for

5. Soft-fused particle likelihood

for $n = 1, \dots, N$ **do**

Combine the two state-conditioned likelihoods:

$$\ell_t^{(n)} \leftarrow \sum_{s \in \{0,1\}} p_t(s) \mathcal{N}(y_t; h_s(x_t^{(n)}), R_s)$$

Update the unnormalized particle weight:

$$\tilde{w}_t^{(n)} \leftarrow w_{t-1}^{(n)} \ell_t^{(n)}$$

end for

6. Particle-weight normalization

for $n = 1, \dots, N$ **do**

Normalize the particle weight:

$$w_t^{(n)} \leftarrow \frac{\tilde{w}_t^{(n)}}{\sum_{m=1}^N \tilde{w}_t^{(m)}}$$

end for

7. CBT estimation

Compute the posterior weighted mean:

$$\hat{x}_t \leftarrow \sum_{n=1}^N w_t^{(n)} x_t^{(n)}$$

8. Degeneracy detection

Compute the effective sample size:

$$N_{\text{eff}} \leftarrow \left[\sum_{n=1}^N \left(w_t^{(n)} \right)^2 \right]^{-1}$$

9. IGA resampling

if $N_{\text{eff}} < N_{\text{th}}$ **then**

Apply the IGA resampling procedure described in Algorithm 1 to obtain $\{\tilde{x}_t^{(n)}\}_{n=1}^N$

for $n = 1, \dots, N$ **do**

Set $x_t^{(n)} \leftarrow \tilde{x}_t^{(n)}$ and $w_t^{(n)} \leftarrow 1/N$

end for

end if

end for

Step 3: Output

return $\{\hat{x}_t, \mathbf{p}_t\}_{t=1}^T$

A.5 Ethical considerations

No new human participants were recruited and no new physiological data were collected as part of this thesis. The systematic review components of Chapters 2 and 3 were based solely on published literature and therefore did not require ethical approval. The benchmarking and modeling studies in Chapters 3–7 involved secondary analysis of previously collected and de-identified datasets. The original studies were conducted under their respective ethical approvals, as reported in the corresponding publications. The datasets were accessed either through publicly available repositories or under the applicable data-sharing arrangements. The secondary analyses conducted in this thesis complied with the relevant data-access conditions and the research governance requirements of the University of Oxford; no additional ethical approval was required.

References

- [1] J. Kuht and A. D. Farmery, ‘Body temperature and its regulation,’ *Anaesthesia & Intensive Care Medicine*, vol. 15, pp. 273–278, 2014.
- [2] C. L. Lim, C. Byrne and J. K. Lee, ‘Human Thermoregulation and Measurement of Body Temperature in Exercise and Clinical Settings,’ *Annals of the Academy of Medicine, Singapore*, vol. 37, pp. 347–353, 2008.
- [3] Wenger, C. B., ‘The Regulation of Body Temperature,’ in *Medical Physiology*, New York: Little, Brown, 1995, pp. 587–613.
- [4] K. Ammer and F. Ring, *The Thermal Human Body: A Practical Guide to Thermal Imaging*. Singapore: Jenny Stanford Publishing, 2019.
- [5] G. J. Tattersall et al., ‘Coping with Thermal Challenges: Physiological Adaptations to Environmental Temperatures,’ *Comprehensive Physiology*, vol. 2, pp. 2151–2202, 2012.
- [6] A. Clarke and K. P. P. Fraser, ‘Why does metabolism scale with temperature?’ *Functional Ecology*, vol. 18, pp. 243–251, 2004.
- [7] A. S. Weller, ‘Body temperature and its regulation,’ *Anaesthesia & Intensive Care Medicine*, vol. 6, pp. 206–209, 2005.
- [8] I. Campbell, ‘Body temperature and its regulation,’ *Anaesthesia & Intensive Care Medicine*, vol. 9, pp. 259–263, 2008.
- [9] D. J. Casa et al., ‘National Athletic Trainers’ Association Position Statement: Exertional Heat Illnesses,’ *Journal of Athletic Training*, vol. 50, pp. 986–1000, 2015.
- [10] R. Carter et al., ‘Epidemiology of Hospitalizations and Deaths from Heat Illness in Soldiers,’ *Medicine & Science in Sports & Exercise*, vol. 37, pp. 1338–1344, 2005.
- [11] Z. Li, Z. J. McKenna, M. R. Kuennen, F. d. C. Magalhães, C. M. Mermier and F. T. Amorim, ‘The Potential Role of Exercise-Induced Muscle Damage in Exertional Heat Stroke,’ *Sports Medicine*, vol. 51, pp. 863–872, 2021.
- [12] C. S. Westwood, J. L. Fallowfield, S. K. Delves, M. Nunns, H. B. Ogden and J. D. Layden, ‘Individual risk factors associated with exertional heat illness: A systematic review,’ *Experimental Physiology*, vol. 106, pp. 191–199, 2021.
- [13] J. D. Périard, D. DeGroot and O. Jay, ‘Exertional heat stroke in sport and the military: Epidemiology and mitigation,’ *Experimental Physiology*, vol. 107, pp. 1111–1121, 2022.
- [14] N. Bouscaren et al., ‘Thermoregulation and Hydration Dynamics in a 160-km Ultra-Endurance Race in a Tropical Environment: A Field Study on 80 Runners,’ *Sports Medicine*, vol. 56, pp. 805–820, 2025.
- [15] H. M. Binkley, J. Beckett, D. J. Casa, D. M. Kleiner and P. E. Plummer, ‘National Athletic Trainers’ Association Position Statement: Exertional Heat Illnesses,’ *Journal of Athletic Training*, vol. 37, pp. 329–343, 2002.
- [16] B. P. McDermott et al., ‘Acute Whole-Body Cooling for Exercise-Induced Hyperthermia: A Systematic Review,’ *Journal of Athletic Training*, vol. 44, pp. 84–93, 2009.

- [17] W. O. Roberts, L. E. Armstrong, M. N. Sawka, S. W. Yeargin, Y. Heled and F. G. O'Connor, 'ACSM Expert Consensus Statement on Exertional Heat Illness: Recognition, Management, and Return to Activity,' *Current Sports Medicine Reports*, vol. 20, pp. 470–484, 2021.
- [18] M. J. Henderson, C. Grandou, B. C. R. Christmas, A. J. Coutts, F. M. Impellizzeri and L. Taylor, 'Core Body Temperatures in Intermittent Sports: A Systematic Review,' *Sports Medicine*, vol. 53, pp. 2147–2170, 2023.
- [19] S. Racinais et al., 'Consensus recommendations on training and competing in the heat,' *Scandinavian Journal of Medicine & Science in Sports*, vol. 25, pp. 6–19, 2015.
- [20] D. S. Moran and L. Mendal, 'Core Temperature Measurement: Methods and Current Insights,' *Sports Medicine*, vol. 32, pp. 879–885, 2002.
- [21] C. Brimicombe et al., 'Wet Bulb Globe Temperature: Indicating Extreme Heat Risk on a Global Grid,' *GeoHealth*, vol. 7, e2022GH000701, 2023.
- [22] G. M. Budd, 'Wet-bulb globe temperature (WBGT)—its history and its limitations,' *Journal of Science and Medicine in Sport*, Heat Stress in Sport, vol. 11, pp. 20–32, 2008.
- [23] H. Guyer, M. Georgescu, D. M. Hondula, F. Wardenaar and J. Vanos, 'Identifying the need for locally-observed wet bulb globe temperature across outdoor athletic venues for current and future climates in a desert environment,' *Environmental Research Letters*, vol. 16, p. 124042, 2021.
- [24] A. Grundstein, E. Cooper and Y. Hosokawa, 'Wet bulb globe temperature variability and its implications on heat stress monitoring,' *International Journal of Sports Science & Coaching*, vol. 19, pp. 1196–1206, 2024.
- [25] H. Yamaguchi et al., 'Using wet-bulb globe temperature meters to examine the effect of heat on various tennis court surfaces,' *Scientific Reports*, vol. 14, p. 15 548, 2024.
- [26] J. R. Brotherhood, 'Heat stress and strain in exercise and sport,' *Journal of Science and Medicine in Sport*, vol. 11, pp. 6–19, 2008.
- [27] H. Inoue et al., 'Incidence and factor analysis for the heat-related illness on the Tokyo 2020 Olympic and Paralympic Games,' *BMJ Open Sport & Exercise Medicine*, vol. 9, e001467, 2023.
- [28] T. Kakamu, K. Wada, D. R. Smith, S. Endo and T. Fukushima, 'Preventing heat illness in the anticipated hot climate of the Tokyo 2020 Summer Olympic Games,' *Environmental Health and Preventive Medicine*, vol. 22, p. 68, 2017.
- [29] M. Callihan et al., 'Comparison of Slate Safety Wearable Device to Ingestible Pill and Wearable Heart Rate Monitor,' *Sensors (Basel, Switzerland)*, vol. 23, p. 877, 2023.
- [30] P. Fulbrook, 'Core temperature measurement in adults: A literature review,' *Journal of Advanced Nursing*, vol. 18, pp. 1451–1460, 1993.
- [31] H. Hymczak et al., 'Core Temperature Measurement-Principles of Correct Measurement, Problems, and Complications,' *International Journal of Environmental Research and Public Health*, vol. 18, p. 10 606, 2021.
- [32] N. A. Taylor, M. J. Tipton and G. P. Kenny, 'Considerations for the measurement of core, skin and mean body temperatures,' *Journal of Thermal Biology*, vol. 46, pp. 72–101, 2014.

- [33] M. P. Castellani, T. P. Rioux, J. W. Castellani, A. W. Potter, S. R. Notley and X. Xu, 'Finite element model of female thermoregulation with geometry based on medical images,' *Journal of Thermal Biology*, vol. 113, p. 103477, 2023.
- [34] V. Melnikov, V. V. Krzhizhanovskaya, M. H. Lees and P. M. Sloom, 'System dynamics of human body thermal regulation in outdoor environments,' *Building and Environment*, vol. 143, pp. 760–769, 2018.
- [35] Y. Zhao, L. de Almeida e Bueno, D. A. Holdsworth and J. H. M. Bergmann, 'Evaluating the Agreement between Oral, Armpit, and Ear Temperature Readings during Physical Activities in an Outdoor Setting,' *International Journal of Environmental Research and Public Health*, vol. 21, p. 595, 2024.
- [36] D. J. Casa et al., 'Validity of devices that assess body temperature during outdoor exercise in the heat,' *Journal of Athletic Training*, vol. 42, pp. 333–342, 2007.
- [37] M. S. Ganio et al., 'Validity and Reliability of Devices That Assess Body Temperature During Indoor Exercise in the Heat,' *Journal of Athletic Training*, vol. 44, no. 2, pp. 124–135, 2009.
- [38] Y. Zhao and J. H. M. Bergmann, 'Estimating core body temperature from heart rate using a residual-compensated adaptive Kalman filter,' *Biocybernetics and Biomedical Engineering*, vol. 45, pp. 617–629, 2025.
- [39] L. P. J. Teunissen, J. Klewer, A. de Haan, J. J. de Koning and H. a. M. Daanen, 'Non-invasive continuous core temperature measurement by zero heat flux,' *Physiological Measurement*, vol. 32, pp. 559–570, 2011.
- [40] D. Hart et al., 'Non-invasive zero-heat-flux technology compared with traditional core temperature measurements in the emergency department,' *The American Journal of Emergency Medicine*, vol. 38, pp. 2383–2386, 2020.
- [41] H.-C. Gunga, M. Sandsund, R. E. Reinertsen, F. Sattler and J. Koch, 'A non-invasive device to continuously determine heat strain in humans,' *Journal of Thermal Biology*, vol. 33, pp. 297–307, 2008.
- [42] P. O. Astrand and I. Ryhming, 'A nomogram for calculation of aerobic capacity (physical fitness) from pulse rate during sub-maximal work,' *Journal of Applied Physiology*, vol. 7, pp. 218–221, 1954.
- [43] A. P. Gagge and R. R. Gonzalez, 'Mechanisms of Heat Exchange: Biophysics and Physiology,' *Comprehensive Physiology*, vol. 1994, pp. 45–84, 1994.
- [44] J. González-Alonso, C. G. Crandall and J. M. Johnson, 'The cardiovascular challenge of exercising in the heat,' *The Journal of Physiology*, vol. 586, pp. 45–53, 2008.
- [45] J. E. Wingo, M. S. Ganio and K. J. Cureton, 'Cardiovascular Drift During Heat Stress: Implications for Exercise Prescription,' *Exercise and Sport Sciences Reviews*, vol. 40, pp. 88–94, 2012.
- [46] M. J. Buller et al., 'Estimation of human core temperature from sequential heart rate observations,' *Physiological Measurement*, vol. 34, pp. 781–798, 2013.
- [47] J. A. J. Stolwijk, 'A Mathematical Model of Physiological Temperature Regulation in Man,' *NASA-Langley, 1971*, 1971.
- [48] D. Fiala, K. J. Lomas and M. Stohrer, 'A computer model of human thermoregulation for a wide range of environmental conditions: The passive system,' *Journal of Applied Physiology*, vol. 87, pp. 1957–1972, 1999.

- [49] R. Niedermann, E. Wyss, S. Annaheim, A. Psikuta, S. Davey and R. M. Rossi, 'Prediction of human core body temperature using non-invasive measurement methods,' *International Journal of Biometeorology*, vol. 58, pp. 7–15, 2014.
- [50] Y. Zhao and J. Bergmann, 'A Hybrid Core Temperature Estimation Method Integrating Physiological and Environmental Factors,' *IEEE Sensors Letters*, vol. 9, pp. 1–4, 2025.
- [51] *Medical Electrical Equipment—Part 2-56: Particular Requirements for Basic Safety and Essential Performance of Clinical Thermometers for Body Temperature Measurement*, Geneva, Switzerland, 2017.
- [52] M. J. Buller, T. Davey, J. L. Fallowfield, S. J. Montain, R. W. Hoyt and S. K. Delves, 'Estimated and measured core temperature responses to high-intensity warm weather military training: Implications for exertional heat illness risk assessment,' *Physiological Measurement*, vol. 41, p. 065 011, 2020.
- [53] Y. Zhao, W. Li and J. H. M. Bergmann, 'Data-driven estimation of core body temperature during physical activity under heat exposure: A systematic review and standardized evaluation,' *Building and Environment*, vol. 297, 2026, Art. no. 114591.
- [54] J. A. Boulant, 'Role of the Preoptic-Anterior Hypothalamus in Thermoregulation and Fever,' *Clinical Infectious Diseases*, vol. 31, S157–S161, 2000.
- [55] C. B. Saper and B. B. Lowell, 'The hypothalamus,' *Current Biology*, vol. 24, R1111–R1116, 2014.
- [56] M. Sund-Levander, C. Forsberg and L. K. Wahren, 'Normal oral, rectal, tympanic and axillary body temperature in adult men and women: A systematic literature review,' *Scandinavian Journal of Caring Sciences*, vol. 16, pp. 122–128, 2002.
- [57] D. Ogoina, 'Fever, fever patterns and diseases called 'fever' – A review,' *Journal of Infection and Public Health*, vol. 4, pp. 108–124, 2011.
- [58] E. Atkins, 'Fever: Its History, Cause, and Function,' *The Yale Journal of Biology and Medicine*, vol. 55, pp. 283–289, 1982.
- [59] J. Pearce, 'A brief history of the clinical thermometer,' *QJM*, vol. 95, pp. 251–252, 2002.
- [60] E. Grodzinsky and M. Sund Levander, 'History of the Thermometer,' in *Understanding Fever and Body Temperature*, E. Grodzinsky and M. Sund Levander, Eds., Cham: Springer International Publishing, 2020, pp. 23–35.
- [61] B. A. MacRae, S. Annaheim, C. M. Spengler and R. M. Rossi, 'Skin Temperature Measurement Using Contact Thermometry: A Systematic Review of Setup Variables and Their Effects on Measured Values,' *Frontiers in Physiology*, vol. 9, p. 29, 2018.
- [62] G. Machin et al., 'Is current body temperature measurement practice fit-for-purpose?' *Journal of Medical Engineering & Technology*, vol. 45, no. 2, pp. 136–144, 2021.
- [63] K. Wang et al., 'Non-contact infrared thermometers for measuring temperature in children: Primary care diagnostic technology update,' *The British Journal of General Practice*, vol. 64, e681–e683, 2014.
- [64] B. Lahiri, S. Bagavathiappan, T. Jayakumar and J. Philip, 'Medical applications of infrared thermography: A review,' *Infrared Physics & Technology*, vol. 55, pp. 221–235, 2012.

- [65] International Organization for Standardization, *Medical electrical equipment—deployment, implementation and operational guidelines for identifying febrile humans using a screening thermograph*, Technical Report, Geneva, Switzerland, 2017.
- [66] K. Cardwell et al., ‘The effectiveness of non-contact thermal screening as a means of identifying cases of Covid-19: A rapid review of the evidence,’ *Reviews in Medical Virology*, vol. 31, e2192, 2021.
- [67] T. Fletcher, A. Whittam, R. Simpson and G. Machin, ‘Comparison of non-contact infrared skin thermometers,’ *Journal of Medical Engineering & Technology*, vol. 42, no. 2, pp. 65–71, 2018.
- [68] G. B. Dell’Isola, E. Cosentini, L. Canale, G. Ficco and M. Dell’Isola, ‘Noncontact Body Temperature Measurement: Uncertainty Evaluation and Screening Decision Rule to Prevent the Spread of COVID-19,’ *Sensors*, vol. 21, p. 346, 2021.
- [69] D. Moher, A. Liberati, J. Tetzlaff, D. G. Altman and the PRISMA Group*, ‘Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement,’ *Annals of Internal Medicine*, vol. 151, pp. 264–269, 2009.
- [70] E. Flemyng et al., ‘Using Risk of Bias 2 to assess results from randomised controlled trials: Guidance from Cochrane,’ *BMJ Evidence-Based Medicine*, vol. 28, no. 4, pp. 260–266, 2023.
- [71] S. Schueler, G. M. Schuetz and M. Dewey, ‘The Revised QUADAS-2 Tool,’ *Annals of Internal Medicine*, vol. 156, no. 4, p. 323, 2012.
- [72] C. G. Teran, J. Torrez-Llanos, T. E. Teran-Miranda, C. Balderrama, N. S. Shah and P. Villarreal, ‘Clinical accuracy of a non-contact infrared skin thermometer in paediatric practice,’ *Child: Care, Health and Development*, vol. 38, pp. 471–476, 2012.
- [73] S. Costanzo and A. Flores, ‘A Non-Contact Integrated Body-Ambient Temperature Sensors Platform to Contrast COVID-19,’ *Electronics*, vol. 9, p. 1658, 2020.
- [74] N. Kameda, ‘Clinical accuracy of non-contact forehead infrared thermometer and infrared tympanic thermometer in postoperative adult patients: A comparative study,’ *Journal of Perioperative Practice*, vol. 32, pp. 142–148, 2022.
- [75] R. Morán-Navarro et al., ‘Validity of Skin, Oral and Tympanic Temperatures During Exercise in the Heat: Effects of Wind and Sweat,’ *Annals of Biomedical Engineering*, vol. 47, pp. 317–331, 2019.
- [76] A. J. Mah, L. Ghazi Zadeh, M. Khoshnam Tehrani, S. Askari, A. H. Gandjbakhche and B. Shadgan, ‘Studying the Accuracy and Function of Different Thermometry Techniques for Measuring Body Temperature,’ *Biology*, vol. 10, p. 1327, 2021.
- [77] S. J. L. Sullivan et al., ‘Clinical evaluation of non-contact infrared thermometers,’ *Scientific Reports*, vol. 11, p. 22 079, 2021.
- [78] T. Basak, S. Aciksoz, B. Tosun, A. Akyuz and C. Acikel, ‘Comparison of three different thermometers in evaluating the body temperature of healthy young adult individuals: Body temperature of healthy young adults,’ *International Journal of Nursing Practice*, n/a–n/a, 2013.
- [79] Y.-M. Kim, M.-R. Jang, J.-R. Moon, G. Park, Y.-J. An and J.-M. Seo, ‘Clinical Accuracy of Non-Contact Forehead Infrared Thermometer Measurement in Children: An Observational Study,’ *Children*, vol. 9, p. 1389, 2022.

- [80] M. Ajčević et al., ‘A Novel Non-Invasive Thermometer for Continuous Core Body Temperature: Comparison with Tympanic Temperature in an Acute Stroke Clinical Setting,’ *Sensors*, vol. 22, p. 4760, 2022.
- [81] D. K. Ng, C.-H. Chan, R. S. Lee and L. C. Leung, ‘Non-contact infrared thermometry temperature measurement for screening fever in children,’ *Annals of Tropical Paediatrics*, vol. 25, pp. 267–275, 2005.
- [82] S. Sollai et al., ‘Performance of a non-contact infrared thermometer in healthy newborns,’ *BMJ Open*, vol. 6, e008695, 2016.
- [83] G. Hayward et al., ‘Non-contact infrared versus axillary and tympanic thermometers in children attending primary care: A mixed-methods study of accuracy and acceptability,’ *British Journal of General Practice*, vol. 70, e236–e244, 2020.
- [84] E. Chiappini et al., ‘Performance of non-contact infrared thermometer for detecting febrile children in hospital and ambulatory settings,’ *Journal of Clinical Nursing*, vol. 20, pp. 1311–1318, 2011.
- [85] I. Franconi, C. La Cerra, A. R. Marucci, C. Petrucci and L. Lancia, ‘Digital Axillary and Non-Contact Infrared Thermometers for Children,’ *Clinical Nursing Research*, vol. 27, pp. 180–190, 2018.
- [86] F. Huang, C. Magnin and P. Brouqui, ‘Ingestible sensors correlate closely with peripheral temperature measurements in febrile patients,’ *Journal of Infection*, vol. 80, pp. 161–166, 2020.
- [87] T. Holder et al., ‘Clinical accuracy of infrared temperature measurement devices: A comparison against non-invasive core-body temperature,’ *Clinical Medicine*, vol. 23, pp. 157–163, 2023.
- [88] N. Hakan, N. Okumuş, M. Aydın, T. Küçüközkan, N. Tuygun and A. Zenciroğlu, ‘The comparison of temporal temperature measurement method by non-contact infrared thermometer with other body temperature measurement methods,’ *Journal of Dr Behcet Uz Children s Hospital*, 2017.
- [89] H. Kocoglu, S. Goksu, M. Isik, Z. Akturk and Y. A. Bayazit, ‘Infrared tympanic thermometer can accurately measure the body temperature in children in an emergency room setting,’ *International Journal of Pediatric Otorhinolaryngology*, vol. 65, pp. 39–43, 2002.
- [90] C. Mahabala, P. H. Dakappa and A. R. Gupta, ‘A novel method for measuring sublingual temperature using conventional non-contact forehead thermometer,’ *F1000Research*, vol. 11, p. 13, 2022.
- [91] H. Apa, S. Gözmen, Ş. Keskin-Gözmen, F. Aslan, N. Bayram and İ. Devrim, ‘Clinical accuracy of non-contact infrared thermometer from umbilical region in children: A new side,’ *The Turkish Journal of Pediatrics*, vol. 58, pp. 180–186, 2016.
- [92] C. Hershler, T. A. Conine, A. Nunn and M. Hannay, ‘Assessment of an infra-red non-contact sensor for routine skin temperature monitoring: A preliminary study,’ *Journal of Medical Engineering & Technology*, vol. 16, pp. 117–122, 1992.
- [93] A. J. E. Bach, I. B. Stewart, A. E. Disher and J. T. Costello, ‘A Comparison between Conductive and Infrared Devices for Measuring Mean Skin Temperature at Rest, during Exercise in the Heat, and Recovery,’ *PLOS ONE*, vol. 10, M. Buchowski, Ed., e0117907, 2015.

- [94] M. Švantner et al., ‘Statistical Study on Human Temperature Measurement by Infrared Thermography,’ *Sensors*, vol. 22, p. 8395, 2022.
- [95] A. A. Fernandes et al., ‘Validity of inner canthus temperature recorded by infrared thermography as a non-invasive surrogate measure for core temperature at rest, during exercise and recovery,’ *Journal of Thermal Biology*, vol. 62, pp. 50–55, 2016.
- [96] B. Unursaikhan et al., ‘Development of a Novel Vital-Signs-Based Infection Screening Composite-Type Camera With Truncus Motion Removal Algorithm to Detect COVID-19 Within 10 Seconds and Its Clinical Validation,’ *Frontiers in Physiology*, vol. 13, p. 905 931, 2022.
- [97] T. Negishi et al., ‘Contactless Vital Signs Measurement System Using RGB-Thermal Image Sensors and Its Clinical Screening Test on Patients with Seasonal Influenza,’ *Sensors*, vol. 20, p. 2171, 2020.
- [98] Q. Wang, Y. Zhou, P. Ghassemi, D. McBride, J. P. Casamento and T. J. Pfefer, ‘Infrared Thermography for Measuring Elevated Body Temperature: Clinical Accuracy, Calibration, and Evaluation,’ *Sensors*, vol. 22, p. 215, 2021.
- [99] T. Bourlai, R. R. Pryor, J. Suyama, S. E. Reis and D. Hostler, ‘Use of Thermal Imagery for Estimation of Core Body Temperature During Precooling, Exertion, and Recovery in Wildland Firefighter Protective Clothing,’ *Prehospital Emergency Care*, vol. 16, pp. 390–399, 2012.
- [100] C. A. Holm, L. Pahler, M. S. Thiese and R. Handy, ‘Evaluation of physiological strain in hot work areas using thermal imagery,’ *Journal of Thermal Biology*, vol. 61, pp. 8–15, 2016.
- [101] S. P. Fenemor, N. D. Gill, S. T. Sims, C. M. Beaven and M. W. Driller, ‘Validity of a Tympanic Thermometer and Thermal Imaging Camera for Measuring Core and Skin Temperature during Exercise in the Heat,’ *Measurement in Physical Education and Exercise Science*, vol. 24, pp. 49–55, 2020.
- [102] C. J. V. D. Heuvel, S. A. Ferguson, D. Dawson and S. S. Gilbert, ‘Comparison of digital infrared thermal imaging (DITI) with contact thermometry: Pilot data from a sleep research laboratory,’ *Physiological Measurement*, vol. 24, pp. 717–725, 2003.
- [103] P. Y. Chan, A. Tay, D. Chen, S. Vogrin, J. McNeil and I. Hopper, ‘Comparison of cutaneous facial temperature using infrared thermography to standard temperature measurement in the critical care setting,’ *Journal of Clinical Monitoring and Computing*, vol. 36, pp. 1029–1036, 2022.
- [104] M. J. Maley, A. P. Hunt, A. J. Bach, C. M. Eglin and J. T. Costello, ‘Infrared cameras overestimate skin temperature during rewarming from cold exposure,’ *Journal of Thermal Biology*, vol. 91, p. 102 614, 2020.
- [105] P. F. Whiting et al., ‘QUADAS-2: A Revised Tool for the Quality Assessment of Diagnostic Accuracy Studies,’ *Annals of Internal Medicine*, vol. 155, pp. 529–536, 2011.
- [106] Y. Zhao and J. H. M. Bergmann, ‘Non-Contact Infrared Thermometers and Thermal Scanners for Human Body Temperature Monitoring: A Systematic Review,’ *Sensors*, vol. 23, p. 7439, 2023.
- [107] Z. Chen, H. Wang, Y. Wang, H. Lin, X. Zhu and Y. Wang, ‘Use of non-contact infrared thermometers in rehabilitation patients: A randomized controlled study,’ *Journal of International Medical Research*, vol. 49, p. 0 300 060 520 984 617, 2021.

- [108] A. Somboonkaew et al., ‘Temperature-compensated infrared-based low-cost mobile platform module for mass human temperature screening,’ *Applied Optics*, vol. 59, E112, 2020.
- [109] G. Sun, T. Matsui, T. Kirimoto, Y. Yao and S. Abe, ‘Applications of Infrared Thermography for Noncontact and Noninvasive Mass Screening of Febrile International Travelers at Airport Quarantine Stations,’ in *Application of Infrared to Biomedical Sciences*, E. Y. Ng and M. Etehadtavakol, Eds., Singapore: Springer, 2017, pp. 347–358.
- [110] J. Foster, A. B. Lloyd and G. Havenith, ‘Non-contact infrared assessment of human body temperature: The journal Temperature toolbox,’ *Temperature*, vol. 8, pp. 306–319, 2021.
- [111] S. Khan et al., ‘Comparative accuracy testing of non-contact infrared thermometers and temporal artery thermometers in an adult hospital setting,’ *American Journal of Infection Control*, vol. 49, pp. 597–602, 2021.
- [112] J. Rubia-Rubia, A. Arias, A. Sierra and A. Aguirre-Jaime, ‘Measurement of body temperature in adult patients: Comparative study of accuracy, reliability and validity of different devices,’ *International Journal of Nursing Studies*, vol. 48, pp. 872–880, 2011.
- [113] F. M. P. Alper, ‘Calibration method of clinical type infrared thermometers and estimation uncertainty in different blackbody cavities,’ *IEEE Instrumentation & Measurement Magazine*, vol. 27, no. 5, pp. 3–10, 2024.
- [114] Y. Sekiguchi, L. N. Belval, R. L. Stearns and D. J. Casa, ‘MONITORING INTERNAL BODY TEMPERATURE,’ vol. 32, 2019.
- [115] J. L. Robinson, R. F. Seal, D. W. Spady and M. R. Joffres, ‘Comparison of esophageal, rectal, axillary, bladder, tympanic, and pulmonary artery temperatures in children,’ *The Journal of Pediatrics*, vol. 133, pp. 553–556, 1998.
- [116] L. P. J. Teunissen and H. A. M. Daanen, ‘Infrared thermal imaging of the inner canthus of the eye as an estimator of body core temperature,’ *Journal of Medical Engineering & Technology*, vol. 35, no. 3-4, pp. 134–138, 2011.
- [117] Y. G. Yohanis and J. D. Mondol, ‘Annual variations of temperature in a sample of UK dwellings,’ *Applied Energy*, vol. 87, pp. 681–690, 2010.
- [118] R. Rising, A. Keys, E. Ravussin and C. Bogardus, ‘Concomitant interindividual variation in body temperature and metabolic rate,’ *American Journal of Physiology-Endocrinology and Metabolism*, vol. 263, E730–E734, 1992.
- [119] A. V. Nguyen et al., ‘Comparison of 3 Infrared Thermal Detection Systems and Self-Report for Mass Fever Screening,’ *Emerging Infectious Diseases*, vol. 16, pp. 1710–1717, 2010.
- [120] H. Kaciuba-Uscilko and R. Grucza, ‘Gender differences in thermoregulation,’ *Current Opinion in Clinical Nutrition & Metabolic Care*, vol. 4, p. 533, 2001.
- [121] E. B. Neves, A. C. C. Salamunes, R. M. De Oliveira and A. M. W. Stadnik, ‘Effect of body fat and gender on body temperature distribution,’ *Journal of Thermal Biology*, vol. 70, pp. 1–8, 2017.
- [122] J. Waalen and J. N. Buxbaum, ‘Is Older Colder or Colder Older? The Association of Age With Body Temperature in 18,630 Individuals,’ *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, vol. 66A, pp. 487–492, 2011.

- [123] J. Iiyama, K. Matsushita, N. Tanaka and K. Kawahira, 'Effects of single low-temperature sauna bathing in patients with severe motor and intellectual disabilities,' *International Journal of Biometeorology*, vol. 52, pp. 431–437, 2008.
- [124] *Medical electrical equipment—part 2-59: Particular requirements for the basic safety and essential performance of screening thermographs for human febrile temperature screening*, Geneva, Switzerland, 2023.
- [125] E. Y.-K. Ng, 'Is thermal scanner losing its bite in mass screening of fever due to SARS?' *Medical Physics*, vol. 32, pp. 93–97, 2005.
- [126] B. C. Pană et al., 'Real-World Evidence: The Low Validity of Temperature Screening for COVID-19 Triage,' *Frontiers in Public Health*, vol. 9, p. 672 698, 2021.
- [127] R. D. Meade, M. P. Poirier, A. D. Flouris, S. G. Hardcastle and G. P. Kenny, 'Do the Threshold Limit Values for Work in Hot Conditions Adequately Protect Workers,' *Medicine and science in sports and exercise*, vol. 48, pp. 1187–1196, 2016.
- [128] D. Dillane and J. A. G. Balanay, 'Comparison between OSHA-NIOSH Heat Safety Tool app and WBGT monitor to assess heat stress risk in agriculture,' *Journal of Occupational and Environmental Hygiene*, 2020.
- [129] X. P. Garzón-Villalba, C. D. Ashley and T. E. Bernard, 'Benchmarking Heat Index as an occupational exposure limit for heat stress,' *Journal of Occupational and Environmental Hygiene*, 2019.
- [130] B. Li, Y. Yang, R. Yao, H. Liu and Y. Li, 'A simplified thermoregulation model of the human body in warm conditions,' *Applied Ergonomics*, vol. 59, pp. 387–400, 2017.
- [131] Y. Kobayashi and S.-i. Tanabe, 'Development of JOS-2 human thermoregulation model with detailed vascular system,' *Building and Environment*, vol. 66, pp. 1–10, 2013.
- [132] A. Joshi, B. Twidwell, M. Park and K. Rykaczewski, 'Comparative analysis of thermoregulation models to assess heat strain in moderate to extreme heat,' *Journal of Thermal Biology*, vol. 127, p. 104 035, 2025.
- [133] A. B. C. G. Silva, L. C. Wrobel and F. L. B. Ribeiro, 'A thermoregulation model for whole body cooling hypothermia,' *Journal of Thermal Biology*, vol. 78, pp. 122–130, 2018.
- [134] J. Xu, A. Psikuta, J. Li, S. Annaheim and R. M. Rossi, 'Evaluation of the convective heat transfer coefficient of human body and its effect on the human thermoregulation predictions,' *Building and Environment*, vol. 196, p. 107 778, 2021.
- [135] X. Xu, A. J. Karis, M. J. Buller and W. R. Santee, 'Relationship between core temperature, skin temperature, and heat flux during exercise in heat,' *European Journal of Applied Physiology*, vol. 113, pp. 2381–2389, 2013.
- [136] V. L. Richmond et al., 'Insulated skin temperature as a measure of core body temperature for individuals wearing CBRN protective clothing,' *Physiological Measurement*, vol. 34, p. 1531, 2013.
- [137] K.-Y. Seng et al., 'Tracking body core temperature in military thermal environments: An extended Kalman filter approach,' in *2016 IEEE 13th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, IEEE, 2016, pp. 296–299.

- [138] I. H. Rizvi and Udayraj, ‘A modified Kalman filter-based model for core temperature estimation during exercise and recovery with/without personal cooling interventions,’ *Journal of Thermal Biology*, vol. 109, p. 103 307, 2022.
- [139] Y. Zhao and J. HM Bergmann, ‘Core body temperature estimation from heart rate via multi-model Kalman filtering and variance-based fusion,’ *Physiological Measurement*, vol. 46, p. 105 002, 2025.
- [140] J. Egbert et al., ‘Accuracy of an estimated core temperature algorithm for agricultural workers,’ *Archives of Environmental & Occupational Health*, 2022.
- [141] M. A. G. Peggen et al., ‘Validity of the estimated core temperature algorithm during real-world prolonged walking exercise under warm ambient conditions,’ *Journal of Thermal Biology*, vol. 125, p. 103 982, 2024.
- [142] C. Kurosaka, T. Maruyama, S. Yamada, Y. Hachiya, Y. Ueta and T. Higashi, ‘Estimating core body temperature using electrocardiogram signals,’ *PLOS ONE*, vol. 17, e0270626, 2022-6-28.
- [143] N. Nazarian et al., ‘Project Coolbit: Can your watch predict heat stress and thermal comfort sensation?’ *Environmental Research Letters*, vol. 16, p. 034 031, 2021.
- [144] S. C. C. Tan, T. C. K. Tran, C. Y. N. Chiang, J. Pan and I. C. C. Low, ‘External auricle temperature enhances ear-based wearable accuracy during physiological strain monitoring in the heat,’ *Scientific Reports*, vol. 14, p. 12 418, 2024.
- [145] S. Laxminarayan et al., ‘Individualized estimation of human core body temperature using noninvasive measurements,’ *Journal of Applied Physiology*, vol. 124, pp. 1387–1402, 2018.
- [146] T. Falcone, F. Cordella, V. Molinaro, L. Zollo and S. Del Ferraro, ‘Real-time human core temperature estimation methods and their application in the occupational field: A systematic review,’ *Measurement*, vol. 183, p. 109 776, 2021.
- [147] C. M. Dolson et al., ‘Wearable Sensor Technology to Predict Core Body Temperature: A Systematic Review,’ *Sensors*, vol. 22, p. 7639, 2022.
- [148] T. Falcone, S. Del Ferraro, V. Molinaro, L. Zollo and P. Lenzuni, ‘A real-time biphasic Kalman filter-based model for estimating human core temperature from heart rate measurements for application in the occupational field,’ *Frontiers in Public Health*, vol. 12, 2024.
- [149] X. Han, J. Wu, Z. Hu, C. Li and B. Sun, ‘Forecasting Human Core and Skin Temperatures: A Long-Term Series Approach,’ *Big Data and Cognitive Computing*, vol. 8, p. 197, 2024.
- [150] V. L. Richmond, S. Davey, K. Griggs and G. Havenith, ‘Prediction of Core Body Temperature from Multiple Variables,’ *The Annals of Occupational Hygiene*, vol. 59, pp. 1168–1178, 2015.
- [151] G. Havenith, S. Davey, V. Downie, K. Griggs and V. Richmond, *Dataset for “Prediction of Core Body Temperature from Multiple Variables” and “The Physiological Strain Index Does Not Reliably Identify Individuals at Risk of Reaching a Thermal Tolerance Limit”: PROSPIE project*, Loughborough University, 2024.

- [152] P. Eggenberger, B. A. MacRae, S. Kemp, M. Bürgisser, R. M. Rossi and S. Annaheim, ‘Prediction of Core Body Temperature Based on Skin Temperature, Heat Flux, and Heart Rate Under Different Exercise and Clothing Conditions in the Heat in Young Adult Males,’ *Frontiers in Physiology*, vol. 9, 2018.
- [153] M. J. Page et al., ‘The PRISMA 2020 statement: An updated guideline for reporting systematic reviews,’ 2021.
- [154] N. E. Moyen, R. C. Bapat, B. Tan, L. A. Hunt, O. Jay and T. Mündel, ‘Accuracy of Algorithm to Non-Invasively Predict Core Body Temperature Using the Kenzen Wearable Device,’ *International Journal of Environmental Research and Public Health*, vol. 18, p. 13 126, 2021.
- [155] N. O. Olen’g and J. Reifman, ‘Hybrid approaches to physiologic modeling and prediction,’ in *Biomonitoring for Physiological and Cognitive Performance during Military Operations*, vol. 5797, SPIE, 2005, pp. 193–203.
- [156] A. V. Gribok, M. J. Buller and J. Reifman $\hat{*}$, ‘Individualized Short-Term Core Temperature Prediction in Humans Using Biomathematical Models,’ *IEEE Transactions on Biomedical Engineering*, vol. 55, pp. 1477–1487, 2008.
- [157] T. Hamatani, A. Uchiyama and T. Higashino, ‘Estimating core body temperature based on human thermal model using wearable sensors,’ in *Proceedings of the 30th Annual ACM Symposium on Applied Computing*, Salamanca Spain: ACM, 2015, pp. 521–526.
- [158] M. Tanaka, M. A. Volle, G. R. Brisson and M. Dion, ‘Body temperature and heart rate relationships during submaximal bicycle ergometer exercises,’ *European Journal of Applied Physiology and Occupational Physiology*, vol. 42, pp. 263–270, 1979.
- [159] T. M. Gibson, P. J. Redman and A. J. Belyavin, ‘Prediction of oesophageal temperatures from core temperatures measured at other sites in man,’ *Clinical Physics and Physiological Measurement*, vol. 2, p. 247, 1981.
- [160] B. L. RD Hansen D Amos, ‘Infrared tympanic temperature as a predictor of rectal temperature in warm and hot conditions,’ *Aviation, space, and environmental medicine*, vol. 67, pp. 1048–1052, 1996.
- [161] A. Gribok, T. McKenna and J. Reifman, ‘Regularization of Body Core Temperature Prediction during Physical Activity,’ in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, 2006, pp. 459–463.
- [162] A. Gribok, W. Rumpler, R. Hoyt and M. Buller, ‘Functional regression for data fusion and indirect measurements of physiological variables collected by wearable sensor systems and indirect calorimetry,’ in *2013 IEEE International Conference on Body Sensor Networks*, 2013, pp. 1–6.
- [163] K.-Y. Seng et al., ‘Nonlinear mixed effects modelling for the analysis of longitudinal body core temperature data in healthy volunteers,’ *Physiological Measurement*, vol. 37, p. 485, 2016.
- [164] O. R. Gibson, A. G. Willmott, C. A. James, M. Hayes and N. S. Maxwell, ‘Power relative to body mass best predicts change in core temperature during exercise-heat stress,’ *The Journal of Strength & Conditioning Research*, vol. 31, pp. 403–414, 2017.

- [165] H. Nakada et al., ‘Development of a method for estimating oesophageal temperature by multi-locational temperature measurement inside the external auditory canal,’ *International Journal of Biometeorology*, vol. 61, pp. 1545–1554, 2017.
- [166] I. Tsadok, M. Scheinowitz, S. A. Shpitzer, I. Ketko, Y. Epstein and R. Yanovich, ‘Assessing rectal temperature with a novel non-invasive sensor,’ *Journal of Thermal Biology*, vol. 95, p. 102788, 2021.
- [167] Y. Li, R. Yao, H. Zhang and C. Du, ‘Exploring on the prediction model of human skin temperature and rectal temperature under heat stress,’ *Building and Environment*, vol. 254, p. 111373, 2024.
- [168] S. Wang et al., ‘Multisensing Wearables for Real-Time Monitoring of Sweat Electrolyte Biomarkers During Exercise and Analysis on Their Correlation With Core Body Temperature,’ *IEEE Transactions on Biomedical Circuits and Systems*, vol. 17, pp. 808–817, 2023.
- [169] M. J. Buller, W. J. Tharion, C. M. Duhamel and M. Yokota, ‘Real-time core body temperature estimation from heart rate for first responders wearing different levels of personal protective equipment,’ *Ergonomics*, vol. 58, pp. 1830–1841, 2015.
- [170] A. P. Hunt, M. J. Buller, M. J. Maley, J. T. Costello and I. B. Stewart, ‘Validity of a noninvasive estimation of deep body temperature when wearing personal protective equipment during exercise and recovery,’ *Military Medical Research*, vol. 6, p. 20, 2019.
- [171] T. Falcone, S. Del Ferraro, V. Molinaro, L. Zollo and P. Lenzuni, ‘Estimation of human core temperature from heart rate: A preliminary study for application in occupational field,’ 2022.
- [172] M. Buller, W. Tharion, R. Hoyt and O. Jenkins, ‘Estimation of Human Internal Temperature from Wearable Physiological Sensors,’ *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 24, pp. 1763–1768, 2010.
- [173] A. P. Welles et al., ‘Estimation of core body temperature from skin temperature, heat flux, and heart rate using a Kalman filter,’ *Computers in Biology and Medicine*, vol. 99, pp. 1–6, 2018.
- [174] S. Kim and J.-Y. Lee, ‘Skin sites to predict deep-body temperature while wearing firefighters’ personal protective equipment during periodical changes in air temperature,’ *Ergonomics*, vol. 59, pp. 496–503, 2016.
- [175] X. Han, J. Wu, Z. Hu, C. Li and X. Hu, ‘A practical deep learning model for core temperature prediction of specialized workers in high-temperature environments,’ *Journal of Thermal Biology*, vol. 128, p. 104079, 2025.
- [176] Y. Zhao and J. H. M. Bergmann, ‘Real-time estimation of core body temperature for heat stress monitoring in hot environments using wearable heart rate sensors,’ *Building and Environment*, vol. 293, 2026, Art. no. 114368.
- [177] J. Mochnac, S. Marchevsky and P. Kocan, ‘Bayesian filtering techniques: Kalman and extended Kalman filter basics,’ in *Proceedings of the 19th International Conference Radioelektronika*, Bratislava, Slovakia: IEEE, 2009, pp. 119–122.
- [178] A. D. H. Smith, D. R. Crabtree, J. L. J. Bilzon and N. P. Walsh, ‘The validity of wireless iButtons® and thermistors for human skin temperature measurement,’ *Physiological Measurement*, vol. 31, no. 1, p. 95, 2009.

- [179] J. D. Stone et al., ‘Assessing the Accuracy of Popular Commercial Technologies That Measure Resting Heart Rate and Heart Rate Variability,’ *Frontiers in Sports and Active Living*, vol. 3, 2021.
- [180] R. Avram et al., ‘Real-world heart rate norms in the Health eHeart study,’ *npj Digital Medicine*, vol. 2, no. 1, p. 58, 2019.
- [181] D. T. Lamarche et al., ‘The recommended Threshold Limit Values for heat exposure fail to maintain body core temperature within safe limits in older working adults,’ *Journal of Occupational and Environmental Hygiene*, vol. 14, no. 9, pp. 703–711, 2017.
- [182] N. Ahmad, R. A. R. Ghazilla, N. M. Khairi and V. Kasi, ‘Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications,’ *International Journal of Signal Processing Systems*, pp. 256–262, 2013.
- [183] W. Zhang et al., ‘Continuous Core Body Temperature Monitoring for Heatstroke Alert via a Wearable In-Ear Thermometer,’ *ACS Sensors*, vol. 10, pp. 1440–1449, 2025.
- [184] V. Venugopal et al., ‘Heat stress and inadequate sanitary facilities at workplaces – an occupational health concern for women?’ *Global Health Action*, 2016.
- [185] D. S. Jardine, ‘Heat Illness and Heat Stroke,’ *Pediatrics In Review*, vol. 28, pp. 249–258, 2007.
- [186] R. Gauer and B. K. Meyers, ‘Heat-Related Illnesses,’ *American Family Physician*, vol. 99, pp. 482–489, 2019.
- [187] J. J. Hess, S. Saha and G. Luber, ‘Summertime Acute Heat Illness in U.S. Emergency Departments from 2006 through 2010: Analysis of a Nationally Representative Sample,’ *Environmental Health Perspectives*, vol. 122, pp. 1209–1215, 2014.
- [188] A. Sheila et al., ‘Heat illness and death among workers - United States, 2012–2013.,’ *MMWR. Morbidity and mortality weekly report*, vol. 63, 2014.
- [189] C. Faurie, B. M. Varghese, J. Liu and P. Bi, ‘Association between high temperature and heatwaves with heat-related illnesses: A systematic review and meta-analysis,’ *Science of The Total Environment*, vol. 852, p. 158 332, 2022.
- [190] M. J. Buller, S. K. Delves, A. L. Fogarty and B. J. Veenstra, ‘On the real-time prevention and monitoring of exertional heat illness in military personnel,’ *Journal of Science and Medicine in Sport*, vol. 24, pp. 975–981, 2021.
- [191] X. P. Garzón-Villalba, Y. Wu, C. D. Ashley and T. E. Bernard, ‘Ability to Discriminate Between Sustainable and Unsustainable Heat Stress Exposures—Part 2,’ *Annals of Work Exposures and Health*, vol. 61, pp. 621–632, 2017.
- [192] S. Laxminarayan, M. Buller, W. Tharion and J. Reifman, ‘Human core temperature prediction for heat-injury prevention,’ *IEEE Journal of Biomedical and Health Informatics*, vol. 19, pp. 883–891, 2015.
- [193] K. Tokizawa, T. Shimuta and H. Tsuchimoto, ‘Validity of a wearable core temperature estimation system in heat using patch-type sensors on the chest,’ *Journal of Thermal Biology*, vol. 108, p. 103 294, 2022.
- [194] M. Soehle, H. Dehne, A. Hoeft and S. Zenker, ‘Accuracy of the non-invasive Tcore™ temperature monitoring system to measure body core temperature in abdominal surgery,’ *Journal of Clinical Monitoring and Computing*, vol. 34, pp. 1361–1367, 2020.

- [195] V. D. Hooper and J. O. Andrews, 'Accuracy of Noninvasive Core Temperature Measurement in Acutely Ill Adults: The State of the Science,' *Biological Research For Nursing*, vol. 8, pp. 24–34, 2006.
- [196] M. Fu, W. Weng, W. Chen and N. Luo, 'Review on modeling heat transfer and thermoregulatory responses in human body,' *Journal of Thermal Biology*, vol. 62, pp. 189–200, 2016.
- [197] A. P. Welles, M. J. Buller, D. P. Looney, W. V. Rumpler, A. V. Gribok and R. W. Hoyt, 'Estimation of metabolic energy expenditure from core temperature using a human thermoregulatory model,' *Journal of Thermal Biology*, vol. 72, pp. 44–52, 2018.
- [198] J. Yang, W. Weng, F. Wang and G. Song, 'Integrating a human thermoregulatory model with a clothing model to predict core and skin temperatures,' *Applied Ergonomics*, vol. 61, pp. 168–177, 2017.
- [199] X. Xu and J. Werner, 'A Dynamic Model of the Human/Clothing/Environment-System.,' *APPLIED HUMAN SCIENCE Journal of Physiological Anthropology*, vol. 16, pp. 61–75, 1997.
- [200] M. Yokota et al., 'Applications of Real-Time Thermoregulatory Models to Occupational Heat Stress: Validation With Military and Civilian Field Studies,' *Journal of Strength and Conditioning Research*, vol. 26, S37–S44, Supplement 2 2012.
- [201] N. Kubota, K. Okada and Y. Yamanaka, 'Circadian Phase Assessment of Core Body Temperature Using a Wearable Temperature Sensor Under the Real World,' *Sleep Science*, vol. 18, e246–e252, 2025.
- [202] G. P. Horn, S. Blevins, B. Fernhall and D. L. Smith, 'Core temperature and heart rate response to repeated bouts of firefighting activities,' *Ergonomics*, 2013.
- [203] S. A. Rubin, 'Core temperature regulation of heart rate during exercise in humans,' *Journal of Applied Physiology*, vol. 62, pp. 1997–2002, 1987.
- [204] C. R. Wyss, G. L. Brengelmann, J. M. Johnson, L. B. Rowell and M. Niederberger, 'Control of skin blood flow, sweating, and heart rate: Role of skin vs. core temperature.,' *Journal of Applied Physiology*, vol. 36, pp. 726–733, 1974.
- [205] L. B. Rowell, 'Human cardiovascular adjustments to exercise and thermal stress.,' *Physiological Reviews*, vol. 54, pp. 75–159, 1974.
- [206] J. Karvonen and T. Vuorimaa, 'Heart Rate and Exercise Intensity During Sports Activities,' *Sports Medicine*, vol. 5, pp. 303–311, 1988.
- [207] N. Verdel, T. Podlogar, U. Ciuha, H.-C. Holmberg, T. Debevec and M. Supej, 'Reliability and Validity of the CORE Sensor to Assess Core Body Temperature during Cycling Exercise,' *Sensors*, vol. 21, p. 5932, 2021.
- [208] M. Khodarahmi and V. Maihami, 'A Review on Kalman Filter Models,' *Archives of Computational Methods in Engineering*, vol. 30, pp. 727–747, 2023.
- [209] G. Davide, 'Understanding Bland Altman analysis,' *Biochemia Medica*, vol. 25, pp. 141–151, 2015.
- [210] D. Fuller et al., 'Reliability and Validity of Commercially Available Wearable Devices for Measuring Steps, Energy Expenditure, and Heart Rate: Systematic Review,' *JMIR mHealth and uHealth*, vol. 8, e18694, 2020.

- [211] D. Bandiera et al., ‘Heat-related risk at Paris 2024: A proposal for classification and review of International Federations policies,’ *British Journal of Sports Medicine*, vol. 58, pp. 860–869, 2024.
- [212] Griffin, Letha and Maxwell, Tanya, ‘A Review of Exertional Heat Illnesses (EHI),’ *Sports Medicine Update*, 2024.
- [213] C. N. Hintz and C. R. Butler, ‘Wearable and ingestible technology to evaluate and prevent exertional heat illness: A narrative review,’ *PM&R*, vol. 16, pp. 398–403, 2024.
- [214] C. Hintz, D. M. Presley and C. R. Butler, ‘Heat stroke burden and validity of wearable-derived core temperature estimation during elite military training,’ *The Physician and Sportsmedicine*, 2024.
- [215] A. S. El-Radhi and W. Barry, ‘Thermometry in paediatric practice,’ *Archives of Disease in Childhood*, vol. 91, pp. 351–356, 2006.
- [216] C. Byrne and C. L. Lim, ‘The ingestible telemetric body core temperature sensor: A review of validity and exercise applications,’ *British Journal of Sports Medicine*, vol. 41, pp. 126–133, 2007.
- [217] W. Weitschies, L. Müller, M. Grimm and M. Koziolok, ‘Ingestible devices for studying the gastrointestinal physiology and their application in oral biopharmaceutics,’ *Advanced Drug Delivery Reviews*, vol. 176, p. 113853, 2021.
- [218] Y. Hashimoto, K. Noto, S. Tada and Y. Nishida, ‘Wearable Multimodal Sensor Probe for Monitoring Core Body Temperature, Electrocardiogram, Heart Rate, and Sweat Rate,’ *IEEE Access*, vol. 13, pp. 70769–70778, 2025.
- [219] D. P. Looney et al., ‘Estimating Resting Core Temperature Using Heart Rate,’ 2018.
- [220] M. K. P., ‘Switching Kalman filters,’ *Technical Reports*, 1998.
- [221] Chen, Zhe, ‘Bayesian Filtering: From Kalman Filters to Particle Filters, and Beyond,’ McMaster University, 2003.
- [222] B. Mor, S. Garhwal and A. Kumar, ‘A systematic review of hidden Markov models and their applications,’ *Archives of Computational Methods in Engineering*, vol. 28, pp. 1429–1448, 2021.
- [223] P. Djuric et al., ‘Particle filtering,’ *IEEE Signal Processing Magazine*, vol. 20, pp. 19–38, 2003.
- [224] J. Elfring, E. Torta and R. van de Molengraft, ‘Particle Filters: A Hands-On Tutorial,’ *Sensors*, vol. 21, p. 438, 2021.
- [225] M. Stone, ‘Cross-Validatory Choice and Assessment of Statistical Predictions,’ *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 36, pp. 111–133, 1974.
- [226] T. Akiba, S. Sano, T. Yanase, T. Ohta and M. Koyama, ‘Optuna: A Next-generation Hyperparameter Optimization Framework,’ in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Anchorage AK USA: ACM, 2019, pp. 2623–2631.
- [227] M. N. Cramer, D. Gagnon, O. Laitano and C. G. Crandall, ‘Human temperature regulation under heat stress in health, disease, and injury,’ *Physiological Reviews*, vol. 102, pp. 1907–1989, 2022.

- [228] L. Shang et al., ‘Time and environmental conditions for heat stress recovery based on human physiological responses: Developing a predictive model,’ *Building and Environment*, vol. 267, p. 112 228, 2025.
- [229] C. Kuptamete and N. Aunsri, ‘A review of resampling techniques in particle filtering framework,’ *Measurement*, vol. 193, p. 110 836, 2022.
- [230] M. N. Sawka, S. N. Cheuvront and R. W. Kenefick, ‘High skin temperature and hypohydration impair aerobic performance,’ *Experimental Physiology*, vol. 97, pp. 327–332, 2012.
- [231] B. J. Velasquez, ‘Monitoring of Core Body Temperature During Exertional Heat Illnesses Emergencies,’ *Sports and Exercise Medicine – Open Journal*, vol. 3, e1–e2, 2017.
- [232] E. E. Coris, A. M. Ramirez and D. J. Van Durme, ‘Heat Illness in Athletes,’ *Sports Medicine*, vol. 34, pp. 9–16, 2004.
- [233] S. Robinson, ‘TEMPERATURE REGULATION IN EXERCISE,’ *Pediatrics*, vol. 32, pp. 691–702, 1963.
- [234] A. W. Nichols, ‘Heat-related illness in sports and exercise,’ *Current Reviews in Musculoskeletal Medicine*, vol. 7, pp. 355–365, 2014.
- [235] A. Hirata et al., ‘Body Core Temperature Estimation Using New Compartment Model With Vital Data From Wearable Devices,’ *IEEE Access*, vol. 9, pp. 124 452–124 462, 2021.
- [236] F. Bakhshi Ostadkalayeh, S. Moradi, A. Asadi, A. Moghaddam Nia and S. Taheri, ‘Performance Improvement of LSTM-based Deep Learning Model for Streamflow Forecasting Using Kalman Filtering,’ *Water Resources Management*, vol. 37, pp. 3111–3127, 2023.
- [237] Y. Zhang et al., ‘Vehicle Dynamics Estimator Utilizing LSTM-Ensembled Adaptive Kalman Filter,’ *IEEE Transactions on Industrial Electronics*, vol. 72, pp. 5429–5439, 2025.
- [238] M. Wang, C. Xu, C. Zhou, Y. Gong and B. Qiu, ‘Study on Underwater Target Tracking Technology Based on an LSTM–Kalman Filtering Method,’ *Applied Sciences*, vol. 12, p. 5233, 2022.
- [239] K. Hajian-Tilaki, ‘Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation,’ *Caspian Journal of Internal Medicine*, vol. 4, pp. 627–635, 2013.
- [240] S. Hochreiter and J. Schmidhuber, ‘Long Short-Term Memory,’ *Neural Computation*, vol. 9, pp. 1735–1780, 1997.
- [241] N. Craft and J. B. Schwartz, ‘Effects of age on intrinsic heart rate, heart rate variability, and AV conduction in healthy humans,’ *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 268, no. 4, H1441–H1452, 1995. PMID: 7733345.
- [242] C. M. Blatteis, ‘Age-Dependent Changes in Temperature Regulation – A Mini Review,’ *Gerontology*, vol. 58, pp. 289–295, 2011.
- [243] R. G. Marks, W. E. Hale, M. T. Moore, F. E. May and R. B. Stewart, ‘Resting heart rates in an ambulatory elderly population: An evaluation of age, sex, symptoms, and medication,’ *Gerontology*, vol. 35, no. 4, pp. 210–217, 1989.

- [244] H.-W. Chow and C.-C. Yang, ‘Accuracy of Optical Heart Rate Sensing Technology in Wearable Fitness Trackers for Young and Older Adults: Validation and Comparison Study,’ *JMIR mHealth and uHealth*, vol. 8, e14707, 2020.
- [245] N. Alugubelli, H. Abuissa and A. Roka, ‘Wearable Devices for Remote Monitoring of Heart Rate and Heart Rate Variability—What We Know and What Is Coming,’ *Sensors*, vol. 22, p. 8903, 2022.
- [246] S. Hu et al., ‘Cold and hot step-changes affecting thermal comfort and physiological indicators in winter,’ *Energy and Buildings*, vol. 254, p. 111 587, 2022.
- [247] ‘The heart rate method for estimating metabolic rate: Review and recommendations,’ *Comparative Biochemistry and Physiology Part A: Molecular & Integrative Physiology*, vol. 158, pp. 287–304, 2011.
- [248] J. A. BUNN, J. W. NAVALTA, C. J. FOUNTAINE and J. D. REECE, ‘Current State of Commercial Wearable Technology in Physical Activity Monitoring 2015–2017,’ *International Journal of Exercise Science*, vol. 11, pp. 503–515, 2018.
- [249] Y. Zhao and J. H. Bergmann, ‘Computational and experimental study of thermal behavior in the oral cavity for biosensing applications,’ *Results in Engineering*, vol. 26, p. 105 284, 2025.
- [250] D. W. DeGroot, K. N. Henderson and F. G. O’Connor, ‘Cooling Modality Effectiveness and Mortality Associate With Prehospital Care of Exertional Heat Stroke Casualties,’ *The Journal of Emergency Medicine*, vol. 64, pp. 175–180, 2023.
- [251] A. Bouchama et al., ‘Classic and exertional heatstroke,’ *Nature Reviews Disease Primers*, vol. 8, p. 8, 2022.
- [252] X. Wu, J. E. Brady, H. Rosenberg and G. Li, ‘Emergency Department Visits for Heat Stroke in the United States, 2009 and 2010,’ *Injury Epidemiology*, vol. 1, p. 8, 2014.
- [253] K. Katić, R. Li and W. Zeiler, ‘Thermophysiological models and their applications: A review,’ *Building and Environment*, vol. 106, pp. 286–300, 2016.
- [254] E. H. Wissler, ‘Pennes’ 1948 paper revisited,’ *Journal of Applied Physiology*, vol. 85, pp. 35–41, 1998.
- [255] G. A. P., ‘An effective temperature scale based on a simple model of human physiological regulatory response,’ *ASHRAE Trans.*, vol. 77, pp. 247–262, 1971.
- [256] S. Takada, T. Sakiyama and T. Matsushita, ‘Validity of the two-node model for predicting steady-state skin temperature,’ *Building and Environment*, vol. 46, pp. 597–604, 2011.
- [257] D. Fiala, G. Havenith, P. Bröde, B. Kampmann and G. Jendritzky, ‘UTCI-Fiala multi-node model of human heat transfer and temperature regulation,’ *International Journal of Biometeorology*, vol. 56, pp. 429–441, 2012.
- [258] M. K. Shourav, S. Salsabila, J.-Y. Lee and J. K. Kim, ‘Estimation of core body temperature by near-infrared imaging of vein diameter change in the dorsal hand,’ *Biomedical Optics Express*, vol. 12, p. 4700, 2021.
- [259] D. Pedreschi et al. ‘Open the Black Box Data-Driven Explanation of Black Box Decision Systems (retrieved //),’ pre-published.
- [260] Y. Wu, B. Sicard and S. A. Gadsden, ‘Physics-informed machine learning: A comprehensive review on applications in anomaly detection and condition monitoring,’ *Expert Systems with Applications*, vol. 255, p. 124 678, 2024.

- [261] A. K. Dubey, A. Kumar, I. Segovia Ramírez and F. P. García Márquez, ‘Machine learning and hybrid intelligence for wind energy optimization: A comprehensive state-of-the-art review,’ *Expert Systems with Applications*, vol. 296, p. 128 926, 2026.
- [262] A. Wang, X. Xue, Z. Gao, Z. Zhang, D. Deng and X. Liu, ‘Physics-encoded neural network via multi-scale tree-structured graph representation for assessing cardiovascular hemodynamics,’ *Expert Systems with Applications*, vol. 299, p. 129 975, 2026.
- [263] M. O. Ericson, Å. Bratt, R. Nisell, U. P. Arborelius and J. Ekholm, ‘Power output and work in different muscle groups during ergometer cycling,’ *European Journal of Applied Physiology and Occupational Physiology*, vol. 55, pp. 229–235, 1986.
- [264] X. Junzeng, W. Qi, P. Shizhang and Y. Yanmei, ‘Error of Saturation Vapor Pressure Calculated by Different Formulas and Its Effect on Calculation of Reference Evapotranspiration in High Latitude Cold Region,’ *Procedia Engineering*, vol. 28, pp. 43–48, 2012.
- [265] J. Malchaire, F. R. d’AMBROSIO Alfano and B. I. Palella, ‘Evaluation of the metabolic rate based on the recording of the heart rate,’ *INDUSTRIAL HEALTH*, vol. 55, pp. 219–232, 2017.
- [266] Y. Yu, X. Si, C. Hu and J. Zhang, ‘A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures,’ *Neural Computation*, vol. 31, pp. 1235–1270, 2019.
- [267] A. O. Ige and M. Sibiya, ‘State-of-the-Art in 1D Convolutional Neural Networks: A Survey,’ *IEEE Access*, vol. 12, pp. 144 082–144 105, 2024.
- [268] S. Nosouhian, F. Nosouhian and A. Kazemi Khoshouei. ‘A Review of Recurrent Neural Network Architecture for Sequence Learning: Comparison between LSTM and GRU (retrieved //),’ pre-published.
- [269] D. Biswas, N. Simoes-Capela, C. Van Hoof and N. Van Helleputte, ‘Heart Rate Estimation From Wrist-Worn Photoplethysmography: A Review,’ *IEEE Sensors Journal*, vol. 19, pp. 6560–6570, 2019.
- [270] S. Ismail, U. Akram and I. Siddiqi, ‘Heart rate tracking in photoplethysmography signals affected by motion artifacts: A review,’ *EURASIP Journal on Advances in Signal Processing*, vol. 2021, p. 5, 2021.
- [271] B. Bent, B. A. Goldstein, W. A. Kibbe and J. P. Dunn, ‘Investigating sources of inaccuracy in wearable optical heart rate sensors,’ *npj Digital Medicine*, vol. 3, p. 18, 2020.
- [272] Y. Sun et al., ‘Comparison of wrist skin temperature with mean skin temperature calculated with Hardy and Dubois’s seven-point method while sleeping,’ *Energy and Buildings*, vol. 259, p. 111 894, 2022.
- [273] W. Liu, Z. Lian, Q. Deng and Y. Liu, ‘Evaluation of calculation methods of mean skin temperature for use in thermal comfort study,’ *Building and Environment*, vol. 46, pp. 478–488, 2011.
- [274] C. Huizenga, H. Zhang, E. Arens and D. Wang, ‘Skin and core temperature response to partial- and whole-body heating and cooling,’ *Journal of Thermal Biology*, vol. 29, pp. 549–558, 2004.
- [275] S. Somvanshi, S. A. Javed, M. M. Islam, D. Pandit and S. Das, ‘A Survey on Kolmogorov-Arnold Network,’ *ACM Computing Surveys*, vol. 58, pp. 1–35, 2026.

- [276] Z. Zhang, Q. Wang, Y. Zhang, T. Shen and W. Zhang, ‘Physics-informed neural networks with hybrid Kolmogorov-Arnold network and augmented Lagrangian function for solving partial differential equations,’ *Scientific Reports*, vol. 15, p. 10 523, 2025.
- [277] B. Zhu and C. Connor, ‘SugarNet: Personalized blood glucose forecast with a Fourier Kolmogorov–Arnold network,’ *Expert Systems with Applications*, vol. 280, p. 127 476, 2025.
- [278] F. Sağlam and Ç. Sözen, ‘A novel Pi-style dendritic neural network based on Kolmogorov–Arnold representation for time series modeling,’ *Expert Systems with Applications*, vol. 297, p. 129 495, 2026.
- [279] P. Ren et al., ‘A Comprehensive Survey of Neural Architecture Search: Challenges and Solutions,’ *ACM Computing Surveys*, vol. 54, pp. 1–34, 2022.
- [280] D. Wen, X. Zhang, X. Liu and J. Lei, ‘Evaluating the consistency of current mainstream wearable devices in health monitoring: A comparison under free-living conditions,’ *Journal of Medical Internet Research*, vol. 19, no. 3, 2017, Art. no. e68.
- [281] G. Tse, ‘Mechanisms of cardiac arrhythmias,’ *Journal of Arrhythmia*, vol. 32, no. 2, pp. 75–81, 2016.
- [282] C. R. Wyndham, ‘Atrial fibrillation: The most common arrhythmia,’ *Texas Heart Institute Journal*, vol. 27, no. 3, pp. 257–267, 2000, PMCID: PMC101077. PMID: 11093410.
- [283] M. Moor et al., ‘Foundation models for generalist medical artificial intelligence,’ *Nature*, vol. 616, pp. 259–265, 2023.
- [284] Y. Luo, Y. Chen, A. Salekin and T. Rahman. ‘Toward Foundation Model for Multivariate Wearable Sensing of Physiological Signals (retrieved //),’ pre-published.
- [285] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang and L. Yang, ‘Physics-informed machine learning,’ *Nature Reviews Physics*, vol. 3, pp. 422–440, 2021.
- [286] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy and A. Bouchachia, ‘A survey on concept drift adaptation,’ *ACM Computing Surveys*, vol. 46, pp. 1–37, 2014.
- [287] S. C. H. Hoi, D. Sahoo, J. Lu and P. Zhao, ‘Online learning: A comprehensive survey,’ *Neurocomputing*, vol. 459, pp. 249–289, 2021.
- [288] O. Anava, E. Hazan, S. Mannor and O. Shamir, ‘Online Learning for Time Series Prediction,’ in *Proceedings of the 26th Annual Conference on Learning Theory*, PMLR, 2013, pp. 172–184.
- [289] J. W. Castellani and A. J. Young, ‘Human physiological responses to cold exposure: Acute responses and acclimatization to prolonged exposure,’ *Autonomic Neuroscience, Thermoregulation*, vol. 196, pp. 63–74, 2016.
- [290] Y. Takahashi et al., ‘Thermoregulation model JOS-3 with new open source code,’ *Energy and Buildings*, vol. 231, p. 110 575, 2021.
- [291] S. Brage et al., ‘Branched equation modeling of simultaneous accelerometry and heart rate monitoring improves estimate of directly measured physical activity energy expenditure,’ *Journal of Applied Physiology*, vol. 96, pp. 343–351, 2004.
- [292] M. Altini, J. Penders and O. Amft, ‘Energy expenditure estimation using wearable sensors: A new methodology for activity-specific models,’ in *Proceedings of the Conference on Wireless Health*, San Diego California: ACM, 2012, pp. 1–8.

- [293] W. Elkhatab et al., ‘Occupational heat-related illness risk screening protocols for aviation: A systematic review,’ *BMJ Public Health*, vol. 3, 2025.
- [294] M. Ahwiadi and W. Wang, ‘An enhanced particle filter technology for battery system state estimation and RUL prediction,’ *Measurement*, vol. 191, 2022, Art. no. 110817.
- [295] S. Schwunk, N. Armbruster, S. Straub, J. Kehl and M. Vetter, ‘Particle filter for state of charge and state of health estimation for lithium–iron phosphate batteries,’ *Journal of Power Sources*, vol. 239, pp. 705–710, 2013.