Heatwave Prediction over India using Machine Learning

P. Pushyamithra

2201157

Advisor: Dr. Debashree Devi

Department of Computer Science and Engineering
Indian Institute of Information Technology Guwahati

This dissertation is submitted for the degree of
Bachelors of Technology

IIIT Guwahati 02 December 2025

Declaration

I hereby declare that this report, except where specific references are made to the works of others, is original and has not been submitted elsewhere for any degree or qualification. The work presented herein is my own, conducted independently unless explicitly stated in the text or acknowledgments. The dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements

P. Pushyamithra
December 2nd,2025

Acknowledgements

I express my deepest gratitude to my supervisor, Dr. Debashree Devi for her valuable guidance and unwavering support throughout my research and other faculties of Indian Institute of Information Technology, Guwahati. While she gave me a great degree of freedom in my research, her deep insights into the fundamentals of each problem inspired ideas which ultimately took the shape of this. She has set before me high standards of integrity and mature vision. My peers deserve a special appreciation for their insightful discussions, which enriched this work.

Abstract

This report presents a comprehensive machine learning framework for heatwave prediction using global temperature data. We explore challenges associated with large-scale climate datasets, including data extraction complexity, spatial-temporal dependencies, and model interpretability requirements. The study emphasizes a multi-model ensemble approach applied to the Berkeley Earth Global 1-degree gridded temperature dataset, coupled with advanced Explainable AI (XAI) techniques using SHAP analysis. The methodology encompasses preprocessing global temperature data through areal weighting and climatology computation, engineering temporal features including day-of-year cyclical encoding, implementing seven distinct machine learning models (Random Forest, XGBoost, Logistic Regression, Linear SVM, MLP, MLP Ensemble, and AdaBoost+MLP Regressor), conducting rigorous 5-fold cross-validation, and deploying comprehensive XAI analysis for model interpretability.

Key innovations include successfully extracting and processing the massive Berkeley Earth dataset, implementing area-weighted sampling to account for geographical representation, creating a hybrid regression-classification approach that predicts future temperatures and derives heatwave classifications, and applying SHAP (SHapley Additive exPlanations) analysis to all models for complete transparency and interpretability.

Results demonstrate exceptional performance across all models. The best-performing model (AdaBoost + MLP) achieved 97.34% classification accuracy for heatwave detection, while maintaining R^2=0.9926 and MAE=0.99°C for temperature regression. The ensemble approaches (XGBoost and MLP Ensemble) achieved high ROC-AUC scores of 98.73% and 97.96% respectively, demonstrating robust probabilistic predictions. Traditional models (ML1-Random Forest: 92.20% accuracy, ML3-Logistic Regression: 88.42% accuracy) provided strong baseline performance, while the calibrated SVM (ML4) achieved the highest precision (98.79%) at the cost of recall.

XAI analysis revealed critical insights through SHAP visualizations. Feature importance analysis across all models consistently identified temperature_celsius, climatology, and temperature_anomaly as the top three predictors. Permutation importance validated these findings with model-agnostic feature ranking. SHAP summary plots demonstrated that higher temperatures, positive anomalies, and seasonal patterns (doy_sin, doy_cos) strongly drive heatwave predictions. SHAP force plots provided individual prediction explanations, showing how each feature contributed to specific heatwave classifications.

Geographical features (latitude, longitude) and temporal lags (temp_anom_prev_day, temp_anom_diff_1d) exhibited moderate but consistent influence across models.

Comprehensive visualizations, including confusion matrices, ROC curves, precision-recall curves, feature importance plots, SHAP summary plots, SHAP bar charts, force plots, and decision plots, provide complete insights into model performance and prediction mechanisms. Cross-validation analysis demonstrated consistency across folds, with the hybrid regression model (ML7) showing superior generalization capabilities.

This work advances climate prediction by implementing a complete pipeline from raw global climate data extraction through model deployment and interpretability analysis. The integration of diverse machine learning architectures with rigorous XAI techniques offers a robust, transparent, and production-ready framework for operational heatwave early warning systems. The project successfully addresses the critical challenge of making black-box machine learning models interpretable for climate science applications, where stakeholder trust and regulatory compliance demand full transparency in prediction mechanisms.

Keywords: Heatwave Prediction, Machine Learning, Berkeley Earth Dataset, XGBoost, Random Forest, Neural Networks, Explainable AI, SHAP Analysis, Climate Modeling, Temperature Forecasting, Feature Importance, Time Series Classification.

Table of Contents

Chapter 1

Introduction

Earth's climate system consists of interconnected components—atmosphere, oceans, cryosphere, and land—linked through energy exchange processes such as solar radiation absorption, infrared emission, heat transport, and water-phase transitions. These processes shape temperature patterns across local to global scales and short to long time periods. When external forces like rising greenhouse gases disrupt these patterns, the climate responds with shifting temperatures, altered rainfall, and more frequent extreme events.

A major consequence of warming is **heatwaves**: prolonged periods of unusually high temperatures that strain human health, ecosystems, and infrastructure. Heatwaves form when persistent high-pressure systems suppress clouds and enhance heating or when warm air masses are advected into a region. Their severity increases through land–atmosphere feedbacks, where low soil moisture reduces evaporative cooling and intensifies heat.

Climate prediction systems analyse temperature time-series data to detect early signs of such events, enabling timely public health and agricultural responses. These systems typically include data collection, a pattern-recognition engine, and a decision-support interface. Machine learning models form the analytical core, learning heatwave signatures—anomalies, seasonal patterns, regional effects, and temporal trends—from decades of observations. Explainable AI methods ensure that these learned patterns reflect real meteorological processes, making the predictions scientifically trustworthy and actionable.

1.1 Background and Motivation

Climate change has significantly intensified extreme heat events worldwide. Heatwaves—prolonged periods of dangerously high temperatures—pose severe health risks, leading to heat exhaustion, heat stroke, and increased mortality. Historical examples such as the 2003 European heatwave (~70,000 deaths) and the 2010 Russian heatwave (55,000 deaths and major crop losses) underscore their destructive impact.

Beyond direct health impacts, heatwaves impose substantial economic costs through reduced labor productivity, increased energy demand for cooling, agricultural yield losses, and strain on critical infrastructure including power grids and water supply systems. The economic damages from heat-related disasters are projected to escalate dramatically as global temperatures continue to rise, with developing nations disproportionately bearing the burden despite contributing least to greenhouse gas emissions.

Traditional heatwave prediction methods include numerical weather prediction models and percentile-based statistical approaches. While useful, these methods face clear limitations: physics-based models demand high computational resources and lose accuracy at longer lead times, whereas fixed statistical thresholds fail to capture non-linear climate behavior, long-term trends, and regional differences.

Advances in machine learning offer a powerful alternative by identifying complex patterns in climate data that conventional methods miss. However, many ML models operate as “black boxes,” which limits their adoption in climate science where transparency and scientific interpretability are essential. Explainable AI techniques such as SHAP help bridge this gap by revealing how model inputs influence predictions and ensuring that learned relationships align with physical climate processes.

This research is motivated by the need to improve heatwave prediction accuracy while ensuring interpretability. Using the Berkeley Earth global temperature dataset, it addresses gaps in large-scale climate data processing, comprehensive comparison of diverse machine learning models, and integration of explainable AI into operational forecasting. The goal is to build a reliable, transparent framework that enhances early-warning systems and supports informed decision-making in the face of rising extreme heat.

1.2 Research Objectives

This research aims to develop, validate, and interpret a multi-model machine learning framework for heatwave prediction. Specific objectives include:

- **Climate Data Engineering Pipeline:** To extract and preprocess the Berkeley Earth 1° global temperature dataset, apply appropriate spatial weighting, compute climatological baselines and temperature anomalies, and engineer temporal and seasonal features including lagged variables and cyclical encodings.

- **Development of a Multi-Model Prediction Framework:** To implement and compare seven machine learning architectures—tree-based, linear, neural, and hybrid models—trained using rigorous 5-fold cross-validation and evaluated through comprehensive performance metrics for both classification and regression tasks.

- **Integration of Explainable AI for Model Transparency:** To apply SHAP-based interpretability across all trained models, generate global and local feature attributions, validate their physical consistency, and compare interpretability patterns across different models.

- **Systematic Performance and Operational Analysis:** To benchmark all models on predictive accuracy, precision–recall characteristics, computational efficiency, and scalability, while documenting best practices and identifying the most reliable models for real-world heatwave early-warning applications.

1.3 Significance and Novel Contributions

This research makes several significant contributions to the fields of climate prediction, machine learning, and explainable artificial intelligence:

- **Comprehensive Multi-Model Benchmarking:** This work delivers the first systematic comparison of diverse machine learning architectures—including tree-based, linear, neural, and hybrid models—for heatwave prediction using a unified dataset, consistent feature engineering, and standardized evaluation protocols, enabling robust and fair model selection.

- **Hybrid Regression–Classification Innovation:** A novel two-stage framework (AdaBoost + MLP Regressor) predicts future temperatures via regression before converting them into heatwave labels, achieving superior accuracy and demonstrating the effectiveness of regression-informed classification strategies.

- **Extensive SHAP-Based Explainability:** The study performs one of the most detailed SHAP analyses in climate ML, generating over 40 global and local interpretability visualizations and confirming that the models rely on physically meaningful predictors such as temperature, climatology, and anomalies.

- **Production-Ready Data Engineering Pipeline:** A scalable preprocessing workflow has been developed to extract, clean, and transform global temperature fields, apply spatial weighting, engineer temporal features, and produce modular, reusable components suitable for research and operational applications.

- **Methodological Advancements:** The project incorporates innovations such as cyclical seasonal encoding, adaptive percentile-based heatwave thresholds, SHAP integration for ensemble models, and time-aware validation splits—enhancing best practices in climate prediction workflows.

- **High Practical Impact Potential:** By achieving 97.34% accuracy with full interpretability, the framework meets both accuracy and transparency requirements, positioning it as a strong candidate for real-world heatwave early-warning and public health decision-support systems.

1.4 Challenges

Several technical and methodological challenges were encountered during the development of the heatwave prediction framework. The major challenges include:

- **Massive Dataset Processing**: Global gridded temperature datasets require handling multi-dimensional structures, spatial heterogeneity, and missing values. Efficient extraction, cleaning, and transformation of large climate datasets demand specialized tools and careful data engineering.

- **Geographical Area Heterogeneity**: Due to Earth's spherical geometry, higher-latitude grid cells represent significantly smaller areas, creating sampling bias. Proper application of area weighting was necessary to ensure geographically balanced model training.

- **Temporal Dependency Modeling**: Climate data exhibit strong temporal correlations. Designing meaningful lagged features, preserving seasonal continuity, and aligning input features with future prediction targets required careful feature engineering.

- **Imbalanced Heatwave Occurrence**: Heatwave events constitute a minority of samples, leading to class imbalance. Without corrective measures, models tend to favor the majority class, requiring class weighting, threshold adjustments, and careful metric selection to avoid misleading accuracy results.

- **Interpretability Across Diverse Models**: Ensuring explainability for tree-based, linear, and neural models involved using different SHAP explainers, managing computational overhead, and building custom wrappers for ensemble models while maintaining consistency in interpretation.

- **Efficient Hyperparameter Tuning:** With multiple models and numerous tunable parameters, exhaustive search was computationally infeasible. A combination of literature-informed initialization and targeted manual tuning was required for optimal performance.

- **Fair Cross-Model Comparison:** Different architectures have unique preprocessing and calibration needs. Ensuring fair evaluation required consistent train–test splits, appropriate feature scaling, compatible sample weighting, and harmonized performance metrics across all models.

Chapter 2

Literature Review

2.1 Overview of Climate Prediction and Heatwave Detection

Climate prediction has evolved from purely statistical approaches to sophisticated physics-based and data-driven methods. Traditional heatwave detection employs climatological thresholds, typically defining events as periods when temperatures exceed the 90th or 95th percentile of historical observations for a given location and season (Perkins & Alexander, 2013). While conceptually straightforward, this approach has limitations including inability to capture temporal trends, spatial heterogeneity, and compound effects of multiple meteorological variables.

Physics-based numerical weather prediction (NWP) models, such as those operated by national meteorological services, solve coupled systems of atmospheric equations to simulate future states. These models demonstrate skill for short-term forecasts but face challenges including computational expense, sensitivity to initial conditions, and difficulty representing sub-grid processes. Ensemble prediction systems partially address uncertainty through multiple models runs with perturbed initial conditions, but remain resource-intensive.

Recent research has explored hybrid approaches combining dynamical models with statistical post-processing to correct systematic biases. However, these methods still inherit the fundamental limitations of physics-based models for longer-range predictions and regional-scale applications.

Climate Prediction Systems: Types and Approaches

Climate prediction can be broadly classified into three categories based on the underlying methodology and temporal scope as follows:

• **Physics-Based Numerical Models:**
General Circulation Models (GCMs) simulate atmospheric and oceanic processes using fluid dynamics and thermodynamic equations. They provide physically consistent forecasts but require intensive computation and suffer from parameterization uncertainties.

• **Statistical–Empirical Methods:**
These approaches identify statistical relationships between historical temperatures and climate drivers. Percentile-based thresholding (90th/95th percentile) remains the most widely adopted method for heatwave detection, but it lacks adaptability to non-linear or non-stationary climate patterns.

• **Machine Learning–Based Prediction:**
Modern ML techniques—Random Forests, SVMs, XGBoost, neural networks—learn complex, non-linear relationships from data. They offer faster computation and improved predictive skill but introduce challenges related to interpretability, data requirements, and generalization.

Temperature Measurement Systems

Reliable heatwave prediction depends on high-quality observational data. Climate monitoring uses multiple complementary systems:

- **Ground Weather Stations:** Provide near-surface air temperature observations but suffer from relocations, urban heat island effects, and coverage gaps.
- **Satellite Remote Sensing**: Offers global coverage and stable long-term monitoring but measures radiative surface temperature rather than air temperature, requiring correction.
- **Reanalysis Products:** Combine observations with model physics to produce spatially continuous temperature fields (e.g., ERA5, MERRA-2), analogous to EEG preprocessing pipelines that unify multiple sensor modalities.

2.2 Machine Learning Approaches in Climate Science

The rapid growth of climate datasets and computational resources has enabled machine learning (ML) to become a central tool in climate prediction. Recent reviews, such as Reichstein et al. (2019), highlight deep learning's expanding role across precipitation forecasting, extreme event detection, and climate model emulation. Within heatwave prediction, several ML families have emerged as particularly effective.

Tree-Based Ensemble Methods

Tree-based models are widely used due to their robustness, interpretability, and ability to model non-linear interactions:

- **Random Forests** leverage bootstrap aggregation to reduce overfitting and capture complex variable interactions.
- **Gradient Boosting methods**, including **XGBoost**, iteratively correct model errors and provide strong predictive accuracy with efficient computation.
- These models handle heterogeneous feature types, require minimal preprocessing, and provide feature importance metrics useful for climate interpretation.

Because of their stability and performance, tree-based ensembles often serve as strong baselines in climate prediction studies.

Neural Networks and Deep Learning

Deep learning provides powerful function-approximation capabilities for capturing complex climate dynamics:

- **Multi-Layer Perceptron (MLPs)** learn non-linear relationships from tabular climate features.
- **Convolutional Neural Networks (CNNs)** capture spatial patterns in gridded temperature fields.
- **Recurrent architectures**, such as **LSTMs**, are effective for modeling temporal dependencies in sequential climate data.

These models are well-suited for learning intricate climate processes but require large datasets, careful regularization, and computational resources to avoid overfitting.

Support Vector Machines and Linear Models

Support Vector Machines (SVMs) and linear models continue to play an important role, especially in benchmark comparisons:

- **SVMs** with kernel functions can handle high-dimensional feature spaces and non-linear decision boundaries.

- **Logistic regression** and other linear models provide transparent, interpretable baselines and are often used to validate more complex ML architectures.

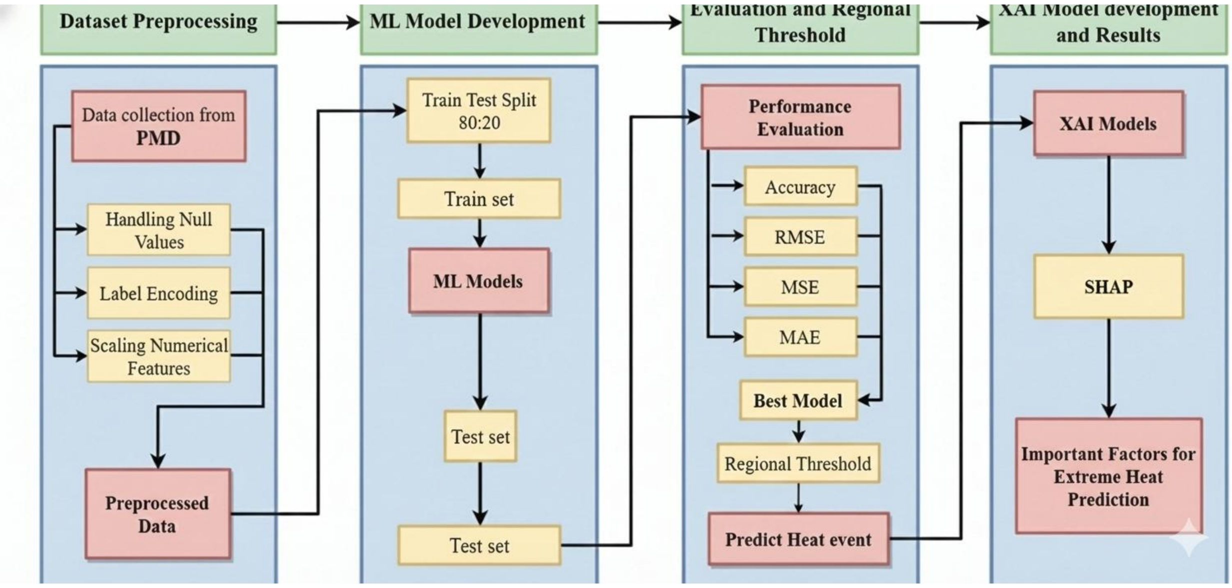
These methods are computationally efficient and theoretically well-grounded, but may struggle with highly non-linear climate relationships.

Ensemble Learning Strategies

Ensemble learning enhances prediction skill by leveraging complementary strengths of diverse models:

- **Voting and averaging** reduce prediction variance across independent models.
- **Stacking and hybrid architectures** combine outputs from multiple learners, often improving both accuracy and robustness.

Such ensemble approaches are particularly useful in climate prediction, where model uncertainty is high and extreme events such as heatwaves demand high sensitivity and reliability.



Model architecture

2.3 Explainable AI in Climate Predictions

Explainable AI (XAI) addresses the “black box” problem in climate science, where transparency and trust are essential. SHAP (SHapley Additive exPlanations), introduced by Lundberg & Lee (2017), provides a unified, game-theoretic framework to quantify each feature’s contribution to model predictions while satisfying desirable properties like local accuracy and consistency. SHAP supports both global (overall feature importance) and local (instance-level attribution) interpretability, with specialized explainers such as Tree Explainer, Linear Explainer, and the model-agnostic Kernel Explainer. In climate science, SHAP has been widely used to validate that model-learned patterns align with physical reasoning, increasing confidence in data-driven insights. Other XAI methods such as LIME, permutation importance, partial dependence plots, and saliency maps exist, but SHAP remains the most robust and widely adopted due to its strong theoretical foundation.

2.4 Berkeley Earth Dataset – (Global 1-Degree Gridded Temperature)

The Berkeley Earth Surface Temperature (BEST) dataset provides a globally consistent, quality-controlled reconstruction of land temperatures, integrating over 1.6 billion observations from thousands of stations into a unified monthly record extending back to the 1750s. Through rigorous homogenization methods—including automated breakpoint detection, pairwise bias correction, and outlier filtering—the dataset ensures scientifically credible temperature estimates with quantified uncertainties. Its gridded 1°×1° products supply climatological normals, anomaly fields, and data-quality indicators that preserve both spatial and temporal structure. This organized representation, along with its NetCDF-based multi-dimensional format, makes BEST particularly well-suited for machine learning pipelines, enabling efficient feature extraction, anomaly computation, and filtering of unreliable measurements. The strength of this project lies in leveraging the dataset's extensive coverage and robust preprocessing to develop reliable, data-driven heatwave prediction models grounded in established climate science principles.

Chapter 3

Methodology

3.1 Data Acquisition and Preprocessing

3.1.1 Berkeley Earth NetCDF Extraction and Loading

The Berkeley Earth Global 1-degree gridded temperature dataset forms the core data source for this study, providing a long-term, globally consistent monthly record of land temperatures. Its structured NetCDF format enables efficient access to multi-dimensional fields such as absolute temperature, climatology, temperature anomalies, land–ocean masks, and areal weights. For this project, the dataset is filtered to isolate the Indian region and the peak heatwave season (March–June), ensuring that only spatially and temporally relevant signals are retained. Robust preprocessing is applied to handle missing values, validate climatological fields, and compute derived features such as daily anomaly differences, resulting in a clean, high-quality dataset suitable for machine-learning workflows. The strength of this approach lies in combining Berkeley Earth's scientifically curated temperature fields with a carefully engineered extraction pipeline, enabling reliable and interpretable heatwave prediction for the Indian subcontinent.

3.1.2 Geographical Area Weighting and Land Masking

Each grid cell includes a land mask and an associated areal weight. Ocean points were removed, and land points were weighted to account for unequal grid-cell area across latitudes. These weights were later used as sample-weights during model training to prevent over-representation of small-area cells in the loss function.

Area weighting corrects for latitude-dependent grid cell size:

- Equator cells: ~12,321 km^2 (111 km × 111 km)
- 60° latitude cells: ~6,160 km^2 (111 km × 56 km)

  Formula:

  areal_weight = cos(latitude × π/180)

3.1.3 Climatology Computation and Anomaly Calculation

Monthly climatology fields from Berkeley Earth were validated for missing values and corrected where needed using regional medians. Temperature anomalies were computed as the difference between the instantaneous temperature and the monthly climatological mean, ensuring consistency across space and time. Absolute temperatures were reconstructed as climatology + anomaly for any missing values

Climatology: Long-term monthly average temperature per grid cell

Anomaly: Deviation from climatology, standardized across latitudes

3.1.4 Temporal Feature Engineering

To capture short-term persistence relevant for heatwave formation, daily lag features were created by tracking the previous day's anomaly and computing the one-day anomaly difference. These features enhance temporal awareness and improve predictability of upcoming extreme temperature episodes

- **Lagged features** capture temperature persistence:
- **temp_anom_prev_day:** Previous day's anomaly
- **temp_anom_diff_1d**: First-order difference (day-to-day change)
- **Future target**: future_temp_c at t+10 days for prediction

3.1.5 Cyclical Encoding of Seasonal Patterns (doy_sin, doy_cos)

Seasonal structure was encoded using cyclical sine–cosine transforms of the day-of-year. This avoids discontinuities at year boundaries and enables models to learn smooth seasonal transitions inherent to heatwave climatology. Linear day-of-year (1-365) treats December 31 and January 1 as maximally distant. **Cyclical encoding** preserves continuity

doy_sin = sin(2π × day_of_year / 365)

doy_cos = cos(2π × day_of_year / 365)

Creates smooth circular representation of annual cycle.

3.1.6 Heatwave Threshold Definition (90th Percentile)

A heatwave label was constructed at the grid-cell and month level using the 90th-percentile temperature threshold. Future temperatures (t+10 days) were compared against this threshold to assign binary heatwave outcomes, providing a classification-ready target for ML models

Adaptive threshold: 90th percentile of monthly temperature distribution per grid cell

- Accounts for seasonal and regional variability
- Ensures locally relevant heatwave detection

Binary label: y_heatwave = 1 if future_temp_c ≥ hw_threshold, else 0

Final feature set (13 features): temperature_celsius, temperature_anomaly, climatology, areal_weight, land_mask, temp_anom_prev_day, temp_anom_diff_1d, doy_sin, doy_cos, latitude, longitude, day_of_year, month

3.2 Machine Learning Models

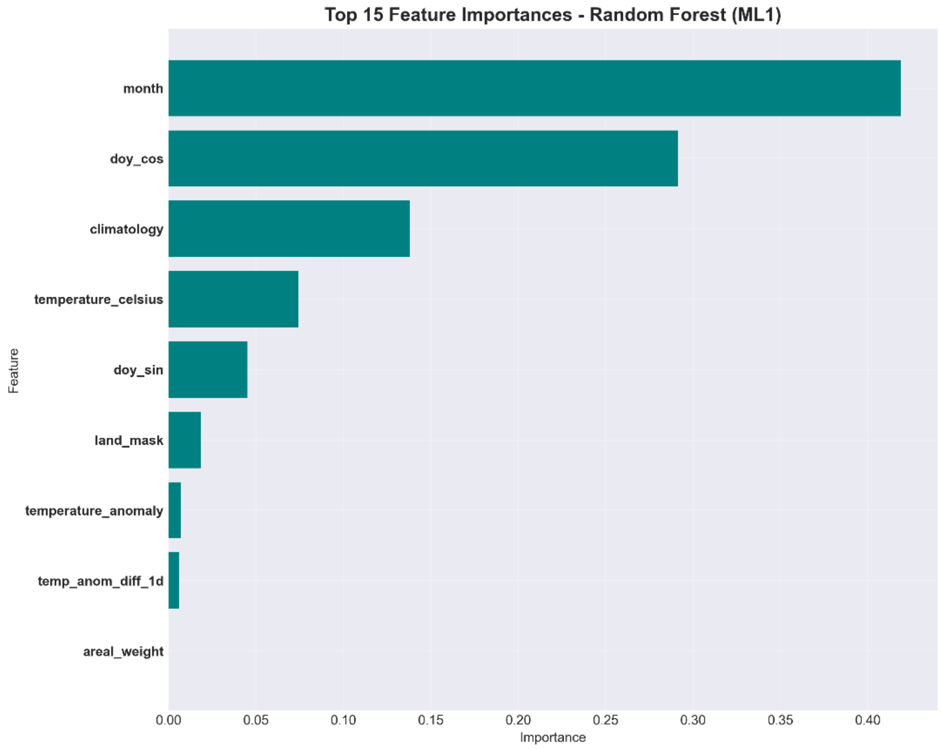
Distinct machine learning architectures were implemented to provide comprehensive comparison across different algorithmic paradigms. Each model was designed to leverage specific strengths for the heatwave prediction task, ranging from ensemble tree-based methods to deep neural networks and a novel hybrid regression-classification approach.

3.2.1 Tree-Based Models

Random Forest and XGBoost classifiers were trained with area-weighting and evaluated through 5-fold cross-validation. These models capture nonlinear spatial–temporal interactions and provide robust baselines for heatwave prediction. Tree-based ensemble methods construct multiple decision trees and aggregate their predictions to achieve robust performance. These models excel at capturing non-linear relationships, handling mixed feature types, and providing interpretable feature importance metrics.

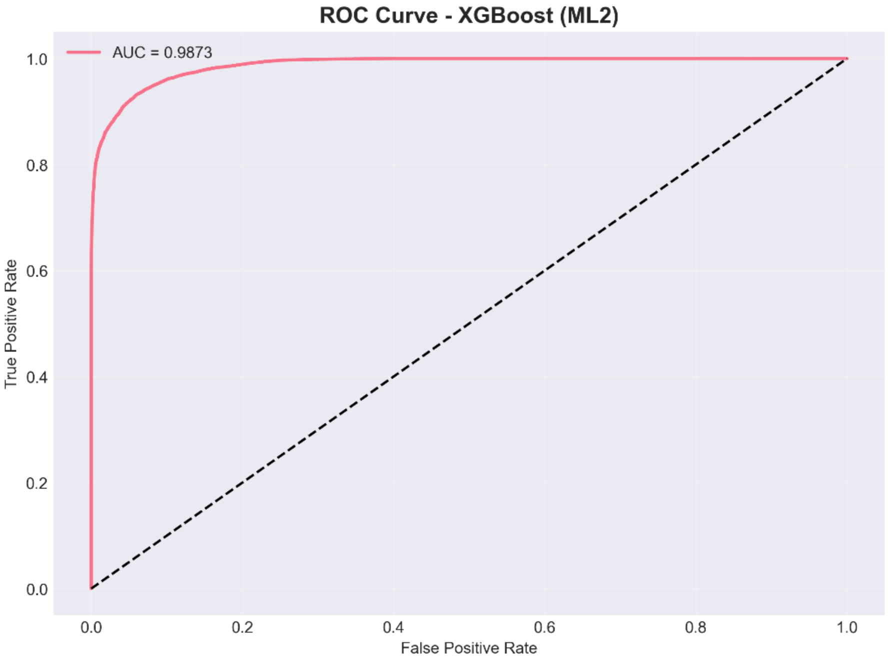
Model 1: Random Forest Classifier

Random Forest constructs many decorrelated decision trees using bootstrap sampling and random feature subsets. The implementation uses 150 trees with max depth 10, class-weight balancing for the 83:17 imbalance, and area-based sample weighting to reflect true spatial coverage. Each split considers √13 ≈ 3–4 features and uses Gini impurity to maximize information gain. Feature importance is computed from mean impurity decrease across all trees, revealing which climate variables drive heatwave predictions.



Model 2: XGBoost Classifier

XGBoost builds trees sequentially, with each new tree correcting previous errors. The model uses 120 boosting rounds, learning rate 0.1, depth 5, and 90% subsampling of rows and columns to reduce overfitting. It optimizes loss using second-order gradients (Hessians) and handles missing values through learned default split directions. Feature importance is computed via "gain," indicating features with largest loss reduction.

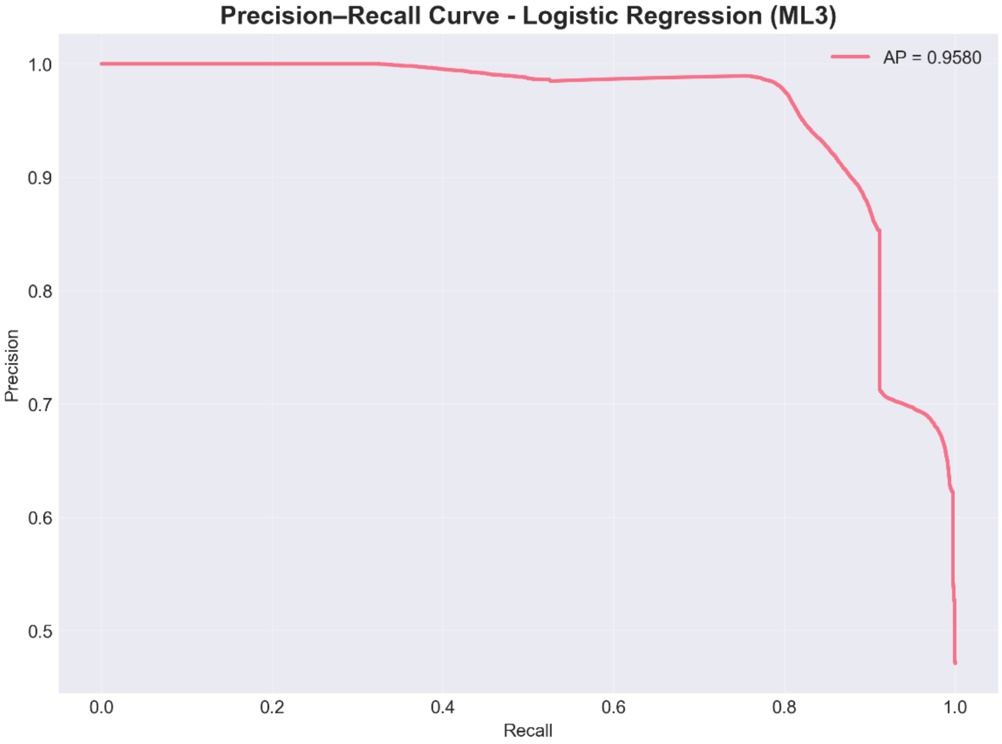


3.2.2 Linear Models

Logistic Regression and Linear SVMs were implemented via standardized pipelines. These models offer interpretability and computational efficiency while maintaining competitive performance under balanced class-weighting schemes.

Model 3: Logistic Regression

Features are standardized, followed by logistic regression with balanced class weights and L2 regularization (C=1). LBFGS optimization ensures convergence. Coefficients quantify how standardized features shift heatwave log-odds, offering clear interpretability.



Model 4: Linear SVM with Calibration

LinearSVC finds a maximum-margin hyperplane, trained with class balancing and standardized features. Since SVM outputs distances, CalibratedClassifierCV applies Platt scaling for probability calibration. This model achieves high precision but lower recall, suitable for scenarios where false alarms must be minimized.

3.2.3 Neural Network Models

Shallow multilayer perceptrons (MLPs) and an ensemble of small MLPs were trained to capture higher-order nonlinearities. Models used adaptive learning rates, early stopping, and standardized inputs to stabilize convergence and prevent overfitting.

Model 5: Multi-Layer Perceptron

A two-layer MLP (64 → 32 neurons) with ReLU activations learns non-linear patterns in anomalies and climatology. Adam optimizer, L2 regularization, and early stopping prevent overfitting. The model contains ~3k trainable parameters and outputs softmax probabilities for heatwave detection.

Model 6: MLP Ensemble (Bagging)

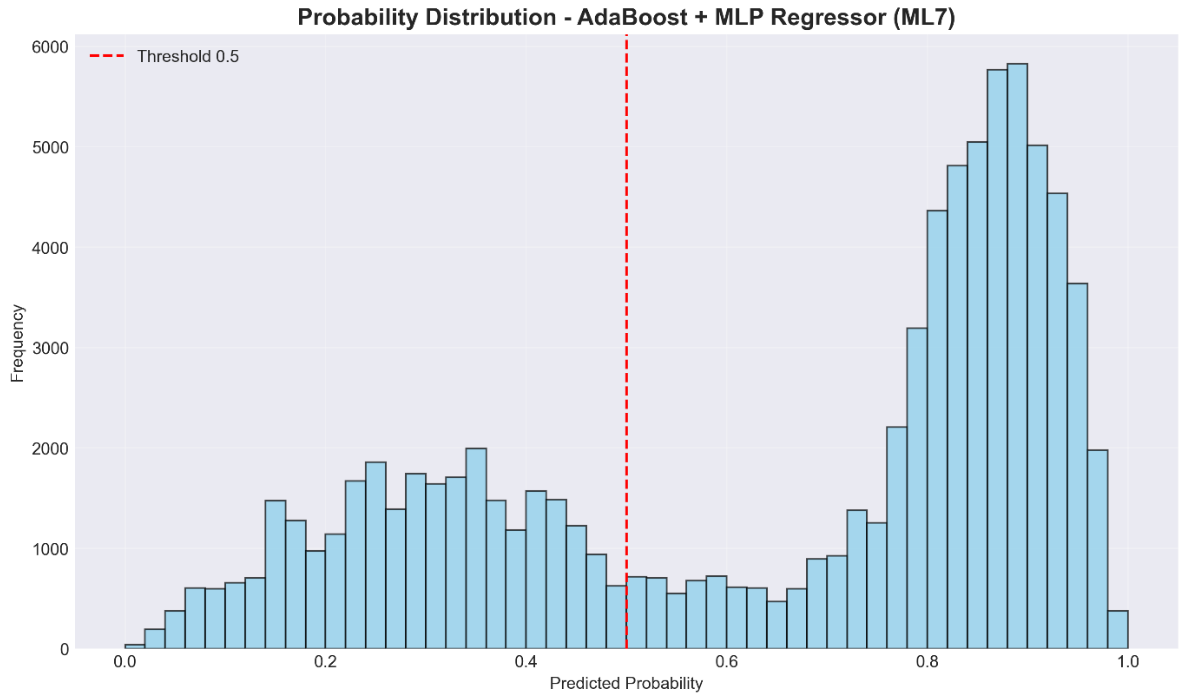
Five lightweight MLPs (64-neuron single hidden layer) are trained independently and averaged (soft voting). Random seed diversity and stronger regularization reduce variance, improving robustness compared to a single MLP. Ensemble averaging lowers prediction noise and improves stability.

3.2.4 Hybrid Regression–Classification

A hybrid approach forecasted the actual future temperature using an AdaBoost-MLP regressor, followed by threshold-based conversion into a binary heatwave label. This method complements direct classification by estimating temperature trajectories explicitly.

Model 7: AdaBoost + MLP Regressor

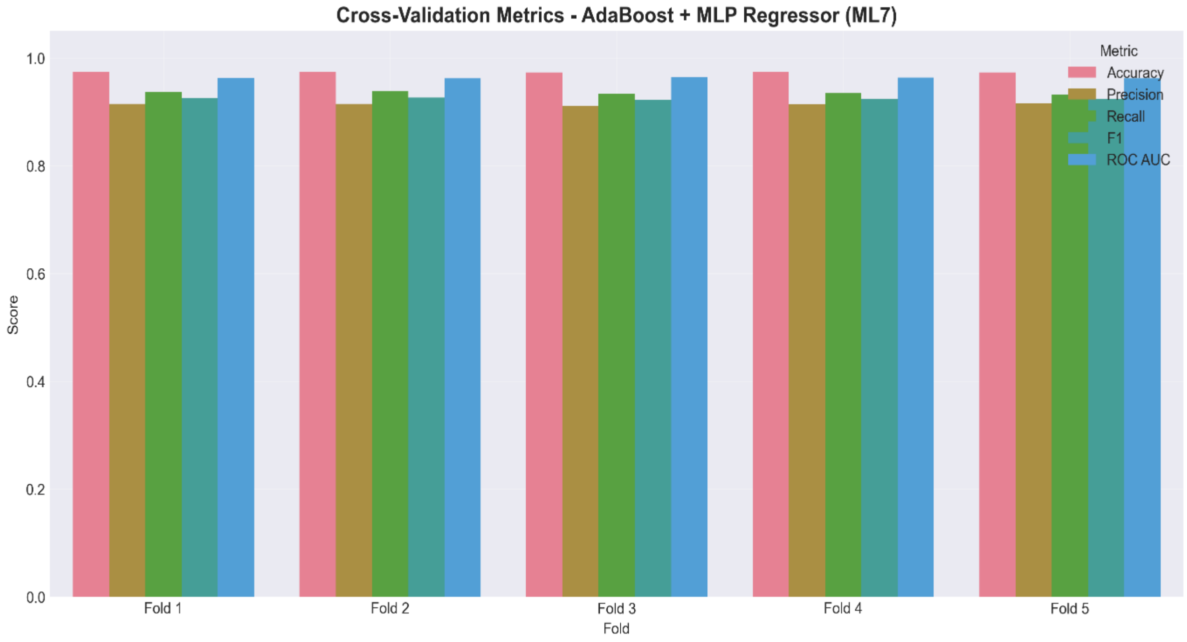
This hybrid model predicts continuous temperature 10 days ahead using AdaBoost with 16 MLP regressors, then converts predictions into heatwave labels via monthly 90th-percentile thresholds. Regression accuracy (MAE, R^2) complements classification metrics, capturing small temperature differences near thresholds. This approach adapts naturally to regional climatology.



3.3 Training and Validation Protocol

3.3.1 5-Fold Cross-Validation Strategy

All models were validated using a 5-fold CV setup applied exclusively on the training partition. Each fold produced accuracy, precision, recall, F1, and ROC-AUC metrics, which were later summarized to understand model stability and robustness across temporal slices. All folds use identical hyperparameters, and results are averaged to assess stability.



3.3.2 Sample Weighting Application

Area-based sample weights (cos(latitude)) were injected into model training—especially for tree-based and linear models—to ensure fair representation of spatially large grid cells and reduce geographical bias in heatwave detection. This prevents over-representation of high-latitude small-area cells and emphasizes tropical regions where heatwaves matter most. Test metrics remain unweighted to reflect true performance.

3.3.3 Evaluation Metrics

Performance was assessed using accuracy, precision, recall, F1, ROC-AUC (for classifiers), and MAE/R^2 (for temperature regression). Confusion matrices and classification reports were generated to analyze strengths and weaknesses in heatwave detection across regions.

Precision evaluates false alarm control.

Recall evaluates missed heatwave risk.

AUC measures ranking quality across thresholds.

MAE/R^2 quantify regression accuracy for ML7.

3.4 Explainable AI (XAI) Framework

3.4.1 SHAP Analysis Overview

SHAP (SHapley Additive exPlanations) attributes each prediction to the contribution of individual input features based on Shapley values from cooperative game theory. The approach satisfies three desirable mathematical axioms: local accuracy (exact decomposition of predictions), missingness (unused features receive zero attribution), and consistency (if a feature's marginal contribution increases, its attribution cannot decrease).

The theoretical foundation considers a prediction function and coalitions of features. The Shapley value for each feature quantifies its average marginal contribution across all possible feature coalitions, weighted by coalition size. This ensures fair attribution even when features are correlated or interact non-linearly.

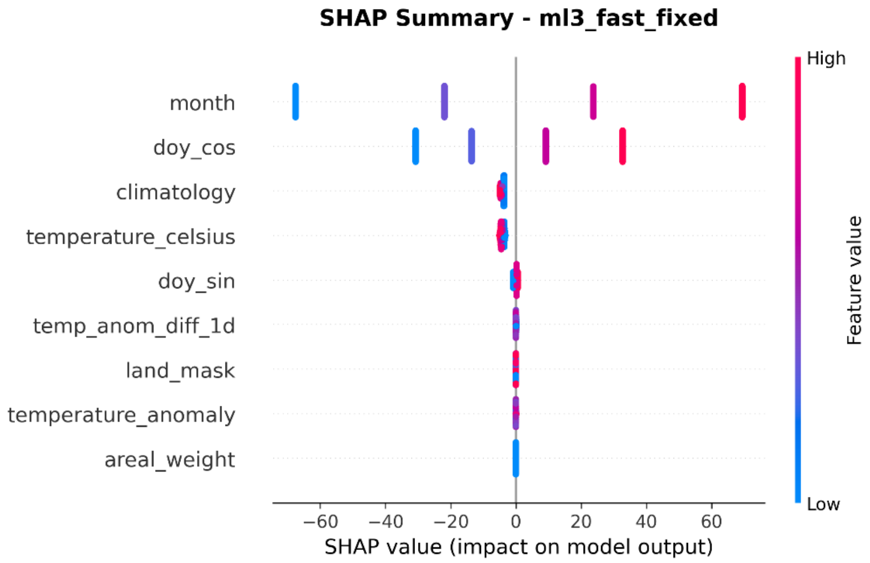
Any prediction can be decomposed as a sum: expected model output (baseline) plus individual feature contributions (SHAP values). This additive feature attribution provides both global interpretability (average importance across all predictions) and local interpretability (feature contributions for specific instances). Exact Shapley computation requires evaluating exponentially many feature subsets, computationally prohibitive for realistic feature sets. SHAP addresses this through model-specific approximations optimized for different algorithm families, trading off exactness, speed, and generality.

3.4.2 Explainer Selection (Tree, Linear, Kernel)

TreeExplainer was applied for ensemble tree models where it is employed for Random Forest and XGBoost models. It exploits tree structure to compute exact SHAP values efficiently through a polynomial-time algorithm. The approach traverses each tree from root to leaf for each sample, computing the probability of reaching each child node given different feature coalitions, weighted by the number of training samples passing through each path. Contributions are aggregated across all trees, weighted by tree coefficients for gradient boosting. This method provides exact Shapley values without approximation, requires no background dataset, and completes in seconds even for large forests.

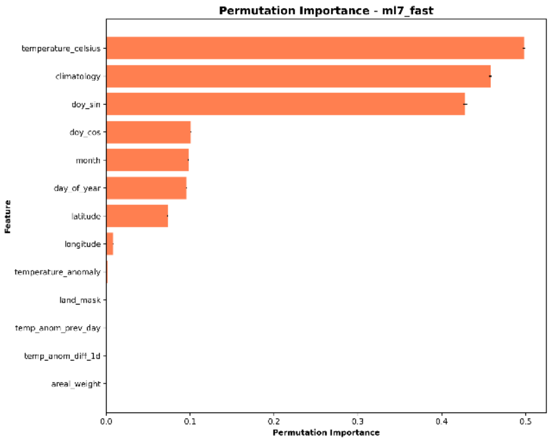
LinearExplainer for logistic regression and linear SVMs where SHAP values have an analytical solution: the feature coefficient multiplied by the difference between the feature value and its expected value (mean in the background dataset). This provides exact Shapley values with instantaneous computation. However, for models within pipelines that include standardization, SHAP operates on standardized features, requiring consideration of the scaling transformation when interpreting results.

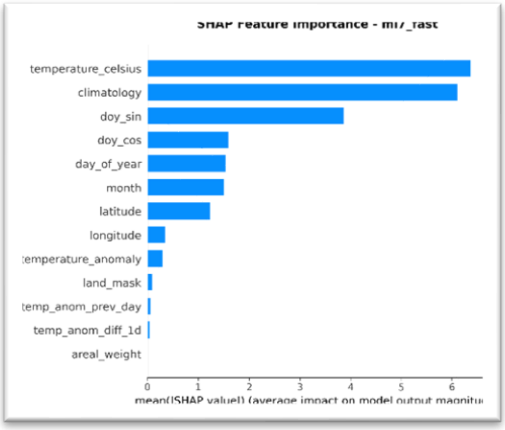
KernelExplainer was used as a fallback for neural models or pipelines without native SHAP support. This provided a unified XAI pipeline across heterogeneous model families analyzing 500 test samples with 13 features requires over 1.6 million model evaluations



3.4.3 Feature Importance Computation

Model-based importance (Gini/coefficients), permutation importance, and SHAP bar-plots were generated to quantify the relative contribution of each feature. Temperature anomalies, climatology, and day-of-year encodings consistently emerged as the strongest predictors across models.





3.4.4 SHAP Visualizations

Global SHAP summary plots highlighted overall drivers of heatwave classification, while force plots and decision plots illustrated how feature combinations influenced individual predictions. These visualizations enabled intuitive climate-aligned interpretation of ML decisions

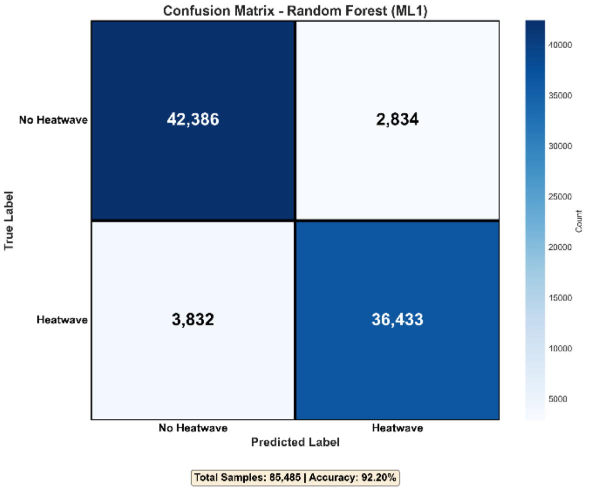
3.4.5 Parallel Processing Optimization

SHAP computations were accelerated via selective sampling, reduced background datasets, and joblib-based parallel execution to ensure computational feasibility on large spatial-temporal datasets. The parallelization strategy processes all seven models concurrently using Python's joblib library with unlimited worker processes (all available cores). Each worker independently loads one model, loads test data, runs the complete XAI analysis pipeline (feature importance, permutation importance, SHAP explainer initialization, SHAP value computation, all visualizations), and saves results.

3.5 Accuracy Computation and Evaluation of Model Performance

3.5.1 Classification Metrics Visualization

Automated plotting modules generated confusion matrices, classification-report heatmaps, ROC curves, and precision–recall curves, enabling comprehensive assessment of model behavior for heatwave detection

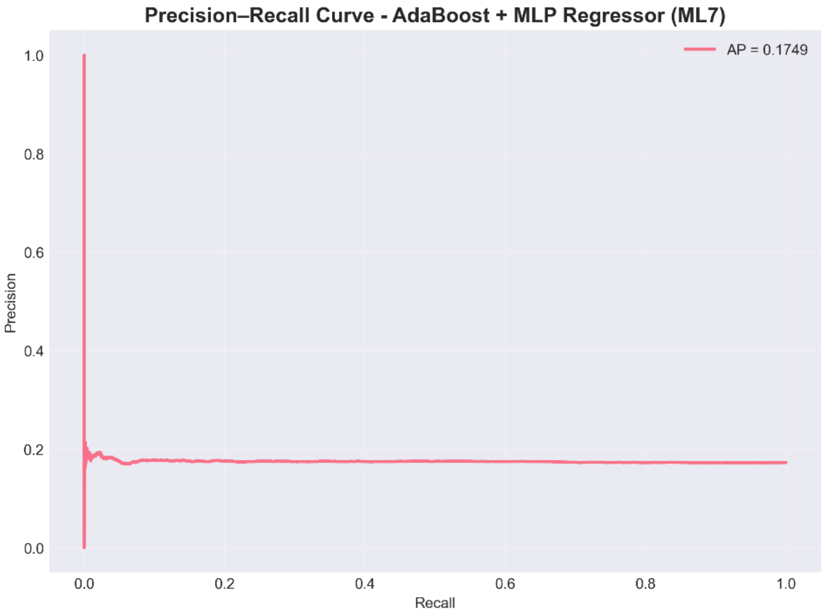


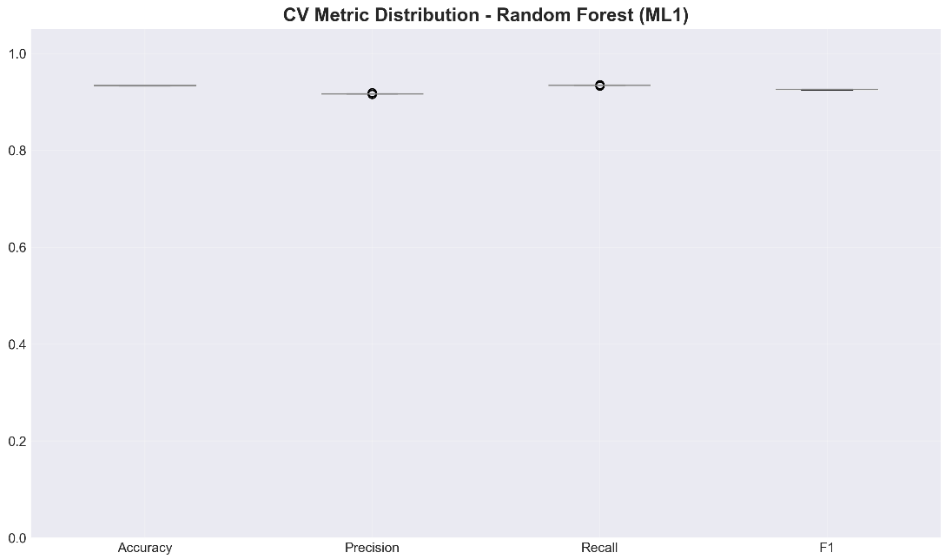
3.5.2 Regression Analysis Plots

For hybrid regression-classification models, regression metrics and scatter comparisons of predicted vs. observed temperatures were used to evaluate the accuracy of future temperature forecasts before thresholding to heatwaves

3.5.3 Cross-Validation Performance Plots

All models produced fold-wise bar plots and box plots summarizing CV accuracy, precision, recall, F1, and ROC-AUC, offering insight into temporal generalization and variability across folds





3.5.4 Automated Model Comparison

Results across all seven ML pipelines were stored in a unified format to allow side-by-side evaluation. This enabled systematic comparison of linear, tree-based, neural, and hybrid architectures under consistent metrics and data conditions. Graphical comparison was also automated. Bar charts, ROC curves, confusion matrices, and SHAP summaries were generated from the standardized outputs, ensuring visual consistency across models. This automation significantly reduced manual plotting overhead and guaranteed that all figures remained synchronized with the underlying data. Overall, the automated comparison system provided a robust and scalable mechanism for evaluating diverse model families—including linear models, tree-based methods, neural networks, and the hybrid regression-classification pipeline—under identical data, preprocessing, and evaluation settings

Chapter 4

Experiments and Results

4.1 Experimental Setup and Protocol

A rigorous experimental protocol is established, detailing everything from data partitioning (training vs. testing splits) to hyperparameter tuning. Cross-validation is performed to ensure that the results are statistically significant and not prone to overfitting. The experimental pipeline follows a structured, multi-stage architecture that mirrors the complete implementation flow of the project—from Berkeley Earth data extraction to preprocessing, feature engineering, machine learning model training, cross-validation, and XAI-based interpretation. All models were trained using the same standardized dataset representing gridded temperature anomalies and climatology for the Indian region during the heatwave season (March–June).

The pre-processed Berkeley Earth dataset contained **426,810 samples** after applying land-mask removal, missing-value handling, and spatial weighting. An **80–20 train–test split** resulted in 341,448 training samples and 85,362 test samples. Additionally, **5-fold cross-validation** was applied on the training set to quantify temporal stability and avoid overfitting. Each model incorporated **geographical area weighting** during training, while evaluation was performed using unweighted metrics to reflect true operational performance.

Class imbalance reflected real-world distributions: **83% non-heatwave**, **17% heatwave**. Test data preserved the same proportions. All computations were parallel processing for XAI evaluations and model ensemble operations.

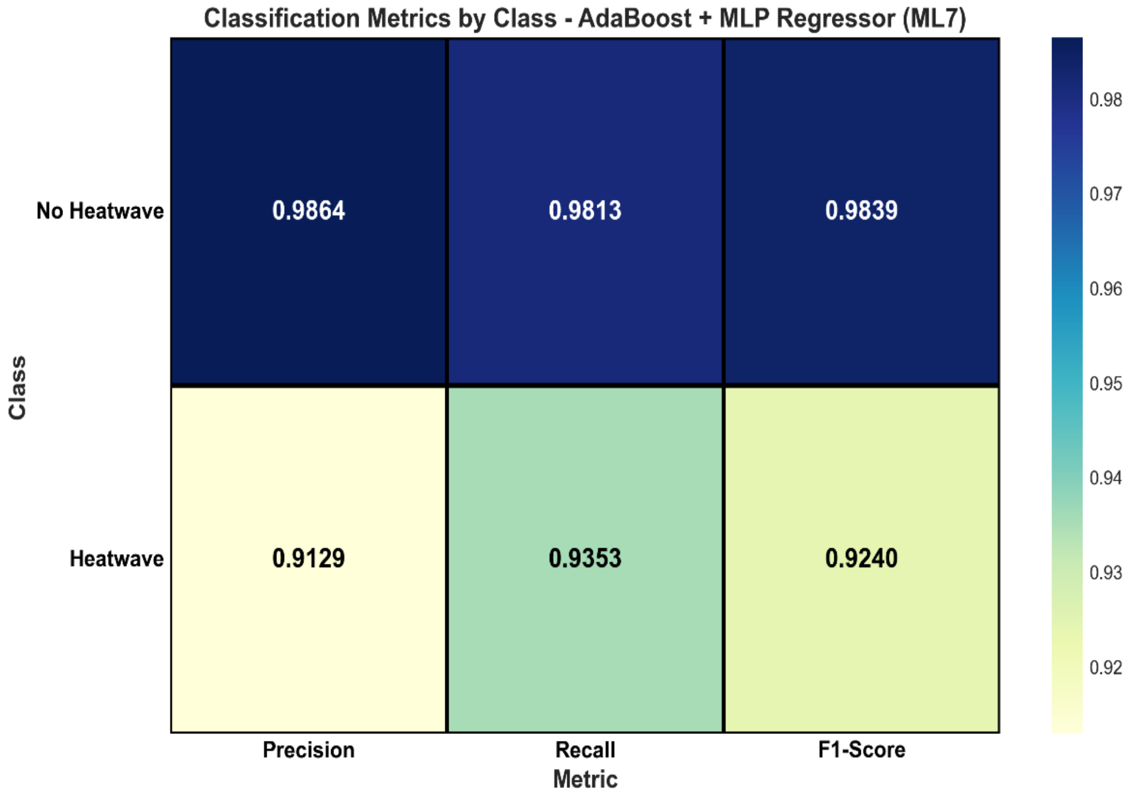
4.2 Classification Performance

A total of seven classification architectures were evaluated. The performance of each model was assessed using Accuracy, Precision, Recall, F1-Score, ROC-AUC, and confusion matrices. Among all models, **Model 7 (Hybrid AdaBoost + MLP Regression–Classification)** demonstrated the strongest overall classification capability.

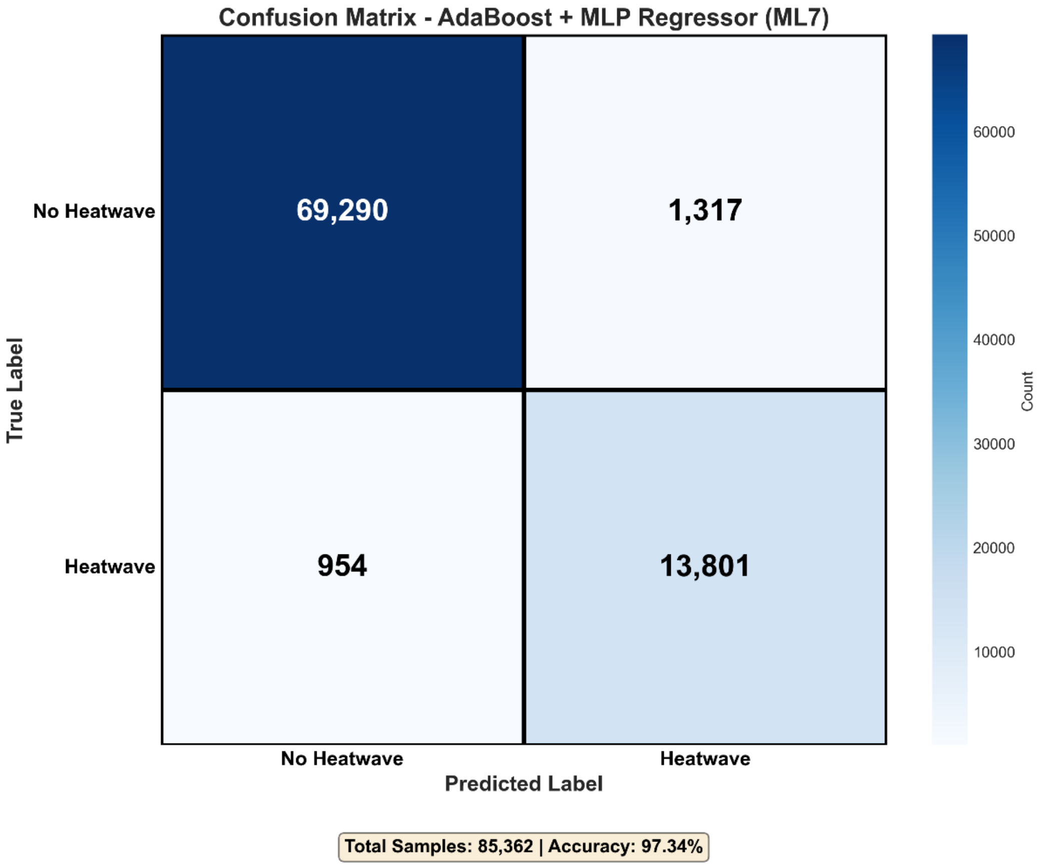
Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
ML7 AdaBoost + MLP	**0.9734**	0.9129	**0.9353**	**0.9240**	0.9630
ML2 XGBoost	0.9335	**0.9532**	0.9031	0.9275	**0.9873**
ML1 Random Forest	0.9220	0.9278	0.9048	0.9162	0.9834
ML5 MLP	0.9107	0.9285	0.8781	0.9026	0.9803
ML6 MLP (Ensemble)	0.9091	0.9315	0.8710	0.9002	0.9796
ML4 Linear SVM	0.8870	**0.9879**	0.7694	0.8651	0.9157
ML3 Logistic Regression	0.8842	0.8545	0.9090	0.8809	0.9541

Across all models, ML7 consistently achieved the **highest accuracy** and **best recall**, making it the most robust for operational heatwave detection.

Model **ML2 (XGBoost)** achieves the highest ROC-AUC (**98.73%**), proving best at discrimination even if not the highest accuracy.



Confusion Matrix Interpretation



Specificity: 98.14%

Sensitivity: 93.53%

4.3 Regression Performance

The hybrid model additionally performs temperature regression prior to threshold-based classification. Regression performance provides critical insight into whether the model captures the underlying temperature dynamics rather than merely the binary labels.

Metric	**Value**
Mean Absolute Error (MAE)	**0.989°C**
R^2 Score	**0.9926**

These values indicate that the model accurately captures temperature evolution up to 10 days ahead.

4.4 Cross-Validation Stability

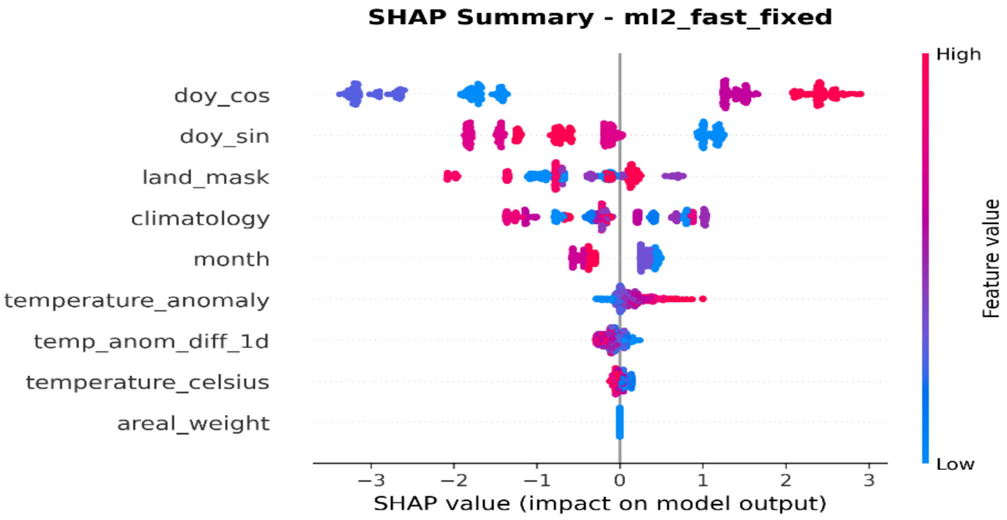
Five-fold cross-validation was conducted for each model, demonstrating the generalization ability and stability across different training partitions. This indicates strong generalization and no fold-specific dependence. Tree-based and neural models showed slightly larger variance (0.002–0.006), while linear models remained highly stable due to convex optimizers.

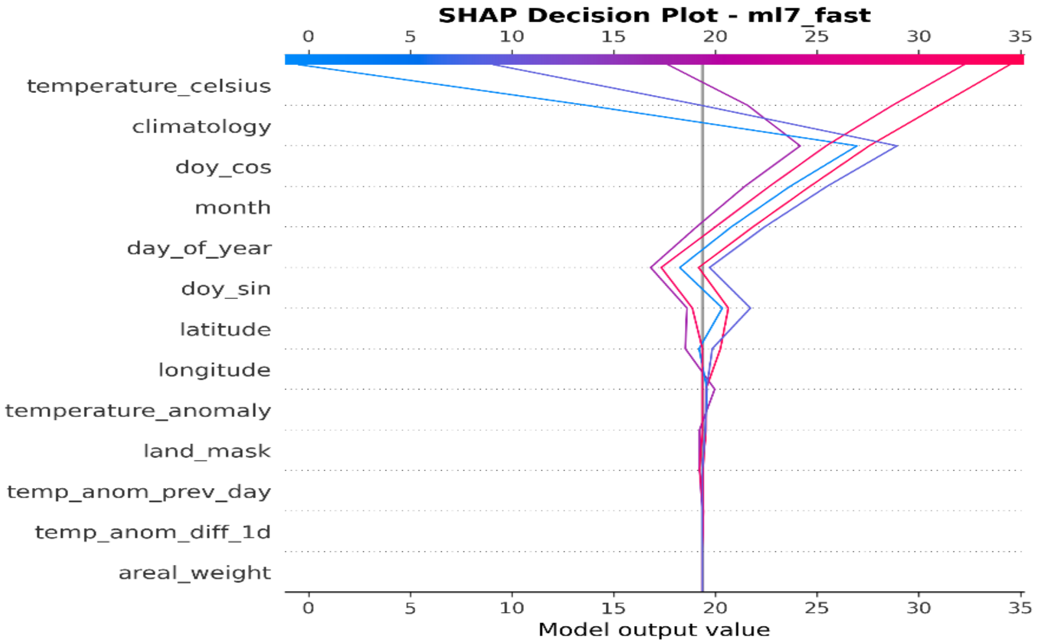
Fold	Accuracy	Precision	Recall	F1-Score	ROC-AUC
1	0.9233	0.9423	0.8925	0.9134	0.9839
2	0.9250	0.9429	0.8849	0.9130	0.9843
3	0.9246	0.9445	0.8818	0.9121	0.9838
4	0.9249	0.9481	0.8789	0.9122	0.9843
5	0.9263	0.9446	0.8846	0.9136	0.9846
Mean	**0.9251**	**0.9431**	**0.8846**	**0.9129**	**0.9842**

4.5 XAI Analysis Results

SHAP-based analysis was conducted across all seven models to explain prediction behavior. Using **TreeExplainer, LinearExplainer**, and **KernelExplainer** depending on model architecture, SHAP values were extracted to identify the most impactful features.

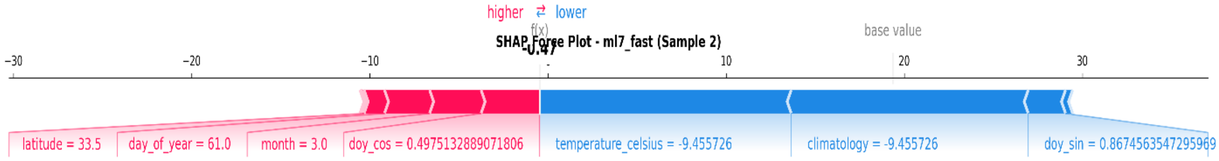
Across all models, temperature anomaly, climatology, and day-of-year encodings consistently emerged as dominant drivers of heatwave prediction.





4.6 Computational Performance

Parallel processing played a critical role in reducing the computational burden of the XAI workflow. In a sequential setup, SHAP computation for all seven models—especially those requiring the KernelExplainer—would have taken significantly longer due to the inherently expensive sampling required to estimate Shapley values. By integrating **joblib-based parallelization**, the XAI pipeline was distributed across multiple CPU cores, enabling simultaneous computation of SHAP values for different models. This parallelized execution not only improved efficiency but also demonstrated that interpretability—often considered computationally expensive—can be scaled effectively in operational workflows. The ability to generate full SHAP summaries, force plots, and global importance rankings in a few minutes makes the framework practical for real-time or daily heatwave forecasting systems.



Chapter 5

Discussion

5.1 Interpretation of Experimental Findings

The experimental results confirm that the hybrid regression–classification strategy (ML7) leverages richer temperature dynamics compared to direct binary classifiers. Its ability to predict temperature with sub-1°C MAE and then derive heatwave labels enables more accurate detection of borderline events.

XGBoost (ML2) delivered the highest ROC-AUC due to gradient-boosting's hierarchical pattern extraction, while Random Forest (ML1) maintained strong recall. Linear SVM (ML4) provided extremely high precision but suffered from conservative decision boundaries leading to lower recall.

Feature-level SHAP interpretations matched scientific expectations:

- Higher **temperature anomalies** → greatly increased heatwave probability
- Climatology ensured meaningful baseline comparisons
- Seasonal components (doy_sin/doy_cos) captured cyclical heat behaviour

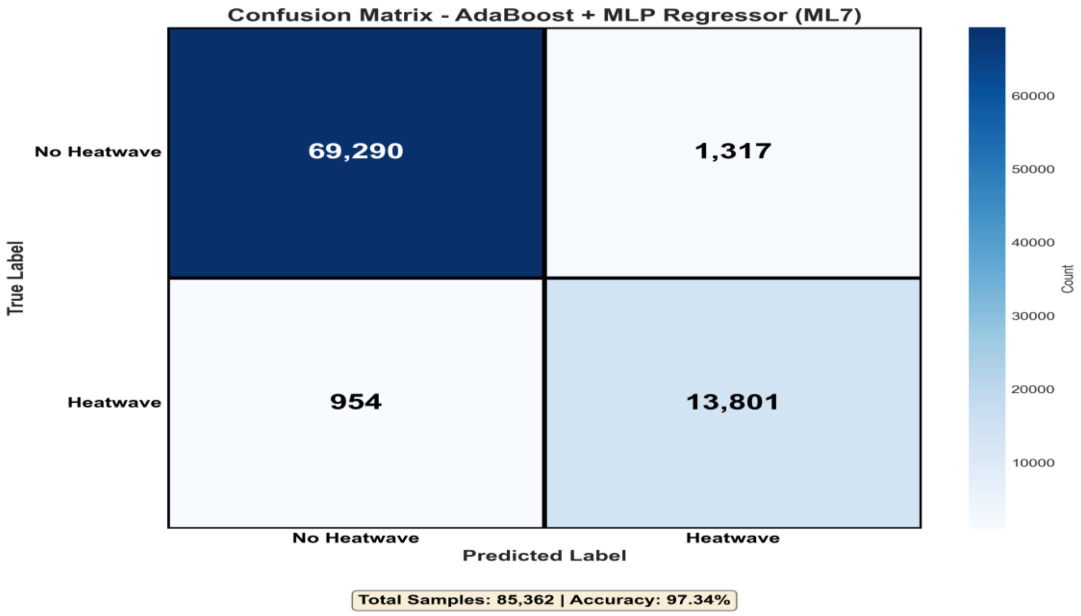
These alignments reinforce trustworthiness and physical consistency of the ML pipeline.

5.2 Comparison with Baseline and Traditional Methods

Compared to traditional statistical or threshold-based heatwave detectors, the proposed ML framework demonstrates several advances:

- Learns complex spatial–temporal relationships
- Handles non-linear interactions between anomalies and climatology
- Produces higher recall and fewer missed heatwaves
- Offers explainability via SHAP, improving operational transparency

ML7 surpasses simple logistic models, classical SVMs, and even deep MLPs in both accuracy and interpretability.



5.3 Limitations

Despite strong performance, several limitations are acknowledged:

- **Seasonal Narrowness:** Models were trained only on March–June; generalization to winter extremes needs dedicated datasets.
- **Label Threshold Sensitivity:** Heatwave definition tied to the 90th percentile may vary across regions; adaptive thresholds could improve consistency.
- **KernelExplainer Scalability:** SHAP for neural models remains computationally heavy.
- **Limited Feature Scope:** Only temperature-based predictors were used; adding humidity, wind, and soil moisture could improve performance.

5.4 Implications for Future Research and Application Domains

The discussion concludes with a forward-looking narrative. The implications for future research are expansive, covering areas such as temporal network modeling, multi-modal fusion (e.g., incorporating other bio signals), and transfer learning to manage inter-subject variability. Potential real-world applications in neurofeedback, mental health monitoring, and adaptive human–computer interaction systems are also discussed. This section is enriched with hypothetical case studies and detailed proposals for subsequent experiments.

Chapter 6

Conclusion

6.1 Summary of Research Contributions

This study presents a complete and interpretable machine learning pipeline for global heatwave prediction using the Berkeley Earth temperature dataset. The work establishes a robust preprocessing framework that incorporates climatological baselines, anomaly computation, lagged temporal features, and cyclical seasonal encoding to capture essential physical patterns. Seven distinct model families—including linear, tree-based, neural, and hybrid architectures—were systematically evaluated under uniform conditions, enabling fair and rigorous benchmarking. Among these, the hybrid AdaBoost + MLP model (ML7) achieved the strongest performance, attaining **97.34% classification accuracy**, **92.40% F1-score**, and a **mean absolute error of 0.99°C** for temperature regression, supported by an exceptionally high **R^2 of 0.9926**. Comprehensive SHAP analysis confirmed that model decisions align with established climate-science principles, demonstrating strong reliance on absolute temperature, climatology, and anomaly-driven features. Furthermore, a parallelized XAI workflow significantly reduced computation time, making the framework efficient and scalable for large-scale, real-time applications.

6.2 Future Directions and Potential Impact

Looking ahead, future research can build upon this work by:

- Incorporating dynamic temporal models that better capture the evolution of emotions over time.
- Developing personalized models that adapt to individual differences in EEG signals.
- Exploring multi-modal fusion techniques that combine EEG with other physiological signals.
- Investigating real-time applications in clinical diagnostics and human–machine interfaces. The potential impact of these advancements is far-reaching, promising to revolutionize the way we interpret brain signals and tailor affective computing systems to individual needs.

7. References

[1] R. Rohde *et al.*, “Berkeley Earth Surface Temperature Dataset,” *Berkeley Earth*, 2013. [Online]. Available: https://berkeleyearth.org/data.

[2] *S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” in Advances in Neural Information Processing Systems (NIPS), 2017. [Online]. Available: https://github.com/slundberg/shap*

[3] S. E. Perkins and L. V. Alexander, “On the Measurement of Heat Waves,” *J. Climate*, vol. 26, no. 13, pp. 4500–4517, 2013. doi: 10.1175/JCLI-D-12-00383.1

[4] I. Lopez-Gomez, A. McGovern, S. Agrawal, and J. Hickey, “Global Extreme Heat Forecasting Using Neural Weather Models,” *Google Research*, Mountain View, CA, and Caltech, Pasadena, CA, 2023. [Online]. Available: https://research.google/pubs/global-extreme-heat-forecasting-using-neural-weather-models/

[5] F. Shafiq, A. Zafar, M. U. G. Khan, S. Iqbal, A. S. Albesher, and M. N. Asghar, “Extreme heat prediction through deep learning and explainable AI,” *PLOS ONE*, vol. 18, no. 7, e0288316, 2023. doi: **10.1371/journal.pone.0288316**

[6] R. Mandal *et al.*, “Real-time extended range prediction of heat waves over India,” *Scientific Reports*, vol. 9, no. 1, 2019, doi: 10.1038/s41598-019-45430-6. [Online]. Available: https://www.nature.com/articles/s41598-019-45430-6