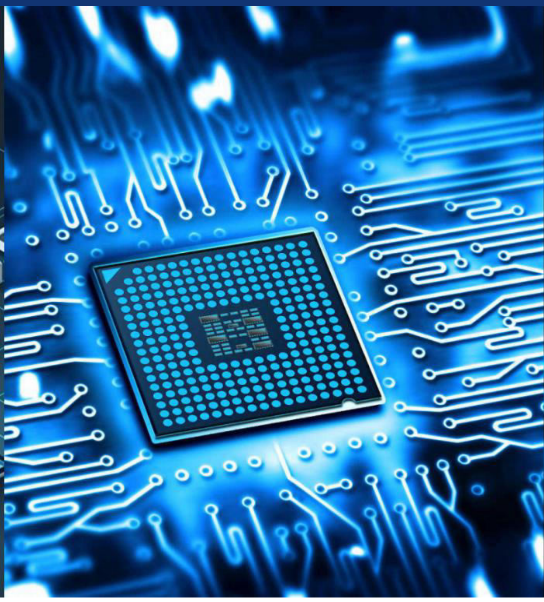
IJIRCCE

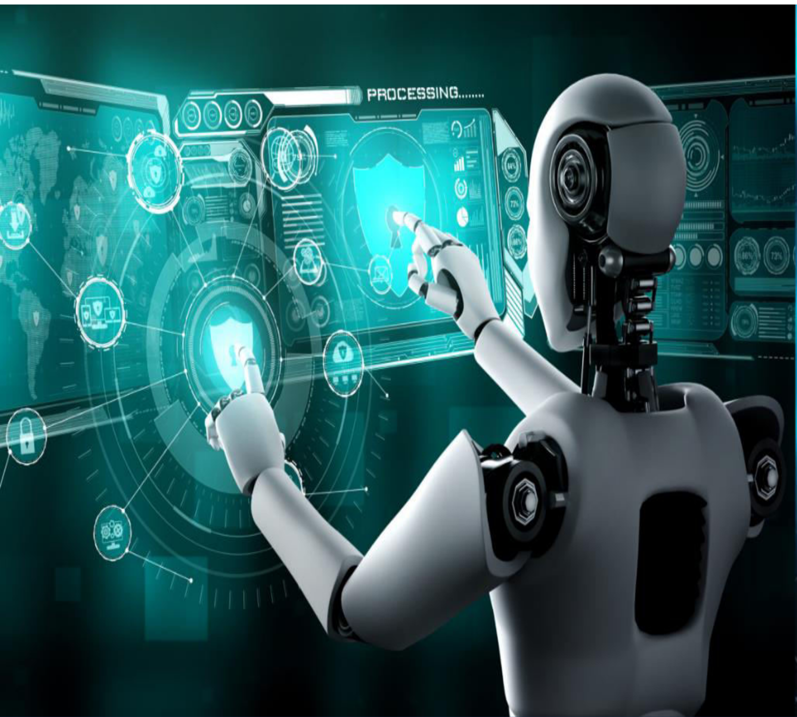
ISSN(O): 2320-9801

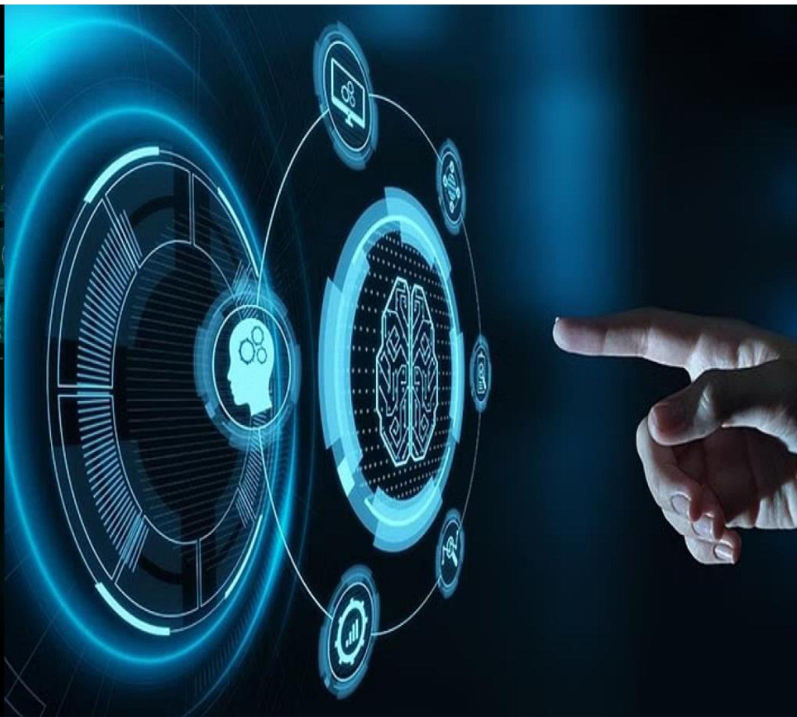
ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Impact Factor: 9.927

Volume 14, Issue 6, June 2026

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Climate Classification Analysis for Chinese Provinces Using Machine Learning Models

Maloba Abraham, Nakalonge Henry Daniel

School of Artificial Intelligence; School of Computer Science and Technology, Nanjing University of Information Science and Technology, Nanjing, China

malobaabraham888@gmail.com; nakalongedaniel@gmail.com

ABSTRACT: This study develops a sophisticated machine learning pipeline to classify climate characteristics across 34 Chinese provinces, leveraging the high-resolution ERA5 reanalysis dataset. The methodology integrates advanced feature engineering, principal component analysis (PCA), clustering, and classification to delineate and predict regional climate patterns. Nineteen enhanced features are derived from meteorological variables, encompassing temperature statistics, humidity variability, pressure dynamics, wind characteristics, and specialized indices such as aridity and temperature-humidity. Robust scaling and PCA are employed to reduce dimensionality while retaining approximately 95% of variance, followed by K-means and hierarchical clustering to group provinces into distinct climate classes. Classification models, including K-nearest neighbors (KNN), random forest, and gradient boosting, are rigorously evaluated for predictive performance. Results demonstrate that K-means clustering with seven clusters achieves the highest silhouette score of 0.350, while gradient boosting yields the highest mean cross-validation accuracy of 0.771 (KNN 0.762, random forest 0.738). Thirteen visualizations, including correlation heatmaps, PCA scatter plots, dendrograms, and radar charts, provide deep insights into feature interactions and climate class profiles. This framework offers a scalable, data-driven solution for regional climate analysis, with significant implications for environmental planning, agricultural optimization, and policy formulation in China's diverse climatic landscape.

KEYWORDS: Climate classification; machine learning; principal component analysis; clustering; ERA5 dataset; Chinese provinces

I. INTRODUCTION

Climate classification serves as a foundational tool for understanding regional environmental patterns, enabling informed decision-making in agriculture, urban development, water resource management, and disaster preparedness. Traditional classification systems, such as the Köppen-Geiger framework [1], rely on predefined thresholds for temperature and precipitation, which often fail to capture the multidimensional nature of climate in geographically diverse regions like China. Spanning arid deserts, temperate plains, and tropical coastlines, China's 34 provinces exhibit complex climatic variability driven by topography, monsoon dynamics, and atmospheric interactions. Machine learning offers a transformative approach to climate classification by integrating diverse meteorological parameters into data-driven models [2], [3], providing nuanced and scalable solutions.

This study proposes a comprehensive machine learning pipeline to classify climate characteristics across China's 34 provinces, utilizing the ERA5 reanalysis dataset from the Copernicus Climate Data Store [4]. The pipeline encompasses advanced feature engineering, dimensionality reduction via PCA, unsupervised clustering, supervised classification, and extensive visualization to identify and predict regional climate patterns. Nineteen features are derived, including statistical summaries of temperature, humidity, pressure, and wind, alongside specialized indices like aridity and temperature-humidity, to create detailed climate profiles. By leveraging robust preprocessing, clustering, and classification techniques, this framework aims to deliver actionable insights for environmental management and policy-making, addressing the limitations of traditional classification systems.

A. Objectives

The primary objective is to develop an enhanced climate classification framework for Chinese provinces using machine learning. Specifically, the study aims to: (1) extract and engineer meaningful climate features from ERA5 meteorological data, (2) apply dimensionality reduction to identify dominant climate drivers, (3) cluster provinces into distinct climate classes based on feature similarities, (4) evaluate the predictive performance of classification models, and (5) generate visualizations to facilitate interpretation and application of results. This framework seeks to provide a scalable, data-driven approach to regional climate analysis, with potential applications in environmental planning, agricultural optimization, and climate resilience strategies.

B. Contributions

This work contributes to climate science by integrating advanced machine learning techniques into a unified pipeline tailored for Chinese provinces. Key contributions include the derivation of 19 enhanced climate features, the application of robust preprocessing to handle outliers, the identification of climate clusters that reflect regional climatic gradients, and the development of predictive models with competitive accuracy. The extensive visualization suite enhances interpretability, making the framework accessible to researchers, policymakers, and practitioners. By leveraging the ERA5 dataset, this study provides a high-resolution, data-driven alternative to traditional climate classification, with implications for regional environmental management.

II. RELATED WORK

Climate classification has evolved significantly since the introduction of rule-based systems like Köppen-Geiger [1], which categorize climates based on temperature and precipitation thresholds. Although these systems remain widely used, and have been refined into updated and high-resolution global maps [5], [6], they struggle to capture the complexity of regions with diverse meteorological and topographic characteristics. Recent advancements in machine learning have enabled more sophisticated approaches. Clustering techniques, such as K-means and hierarchical clustering, have been employed to group regions based on meteorological similarities, revealing patterns not captured by traditional methods; for China specifically, spectral and nearest-neighbor clustering of station data [2] and k-means/k-medoids analyses of major cities [3] have produced regionalizations that outperform threshold-based schemes. Dimensionality reduction techniques, notably PCA, have identified dominant climate features, reducing computational complexity while preserving critical information [7]. Ensemble classifiers, such as random forests, have demonstrated robust performance in predicting climate zones, leveraging multiple features to improve accuracy [8].

The ERA5 dataset, with its high spatial and temporal resolution, has become a cornerstone for climate studies [9], superseding earlier reanalyses such as ERA-Interim [10] that underpinned many regional analyses. Previous research has utilized these reanalyses for regional studies, but few have integrated feature engineering, clustering, and classification into a unified pipeline for Chinese provinces. This work builds on these efforts by incorporating advanced indices (e.g., aridity, temperature-humidity) and robust preprocessing to enhance classification accuracy and interpretability. By addressing the multidimensional nature of climate data, this study offers a novel framework for regional climate analysis, with potential to inform policy and resource management.

III. METHODOLOGY

The methodology employs a multi-step pipeline implemented in Python, utilizing libraries such as xarray, pandas, NumPy [11], scikit-learn [12], seaborn, and matplotlib [13]. The pipeline is designed to maximize the extraction of meaningful climate patterns while ensuring computational efficiency and robustness. It comprises data acquisition, feature engineering, preprocessing, dimensionality reduction, clustering, classification, and visualization, each detailed below.

A. Data Acquisition and Setup

Meteorological data for 34 Chinese provinces are sourced from the ERA5 single-levels dataset, accessed via the Copernicus Climate Data Store [4]. The dataset includes variables such as 2-meter temperature (t2m), dewpoint temperature (d2m), surface pressure (sp), 10-meter wind components (u10, v10), and total precipitation (tp), stored in NetCDF format. The analysis is conducted in a Google Colab environment, with data stored in Google Drive and copied to a local directory for processing. Directories for data inputs (data_dir) and visualization outputs (plots_dir) are created to ensure organized workflow management.

B. Enhanced Feature Engineering

A custom function, enhanced_climate_parameters, processes raw ERA5 data to derive 19 climate features, enhancing the representation of regional climate characteristics. Temperature parameters (mean, maximum, minimum, standard deviation, and range) are calculated by converting Kelvin to Celsius. Relative humidity is computed using the Magnus-Tetens approximation [14], based on temperature and dewpoint data, with mean and standard deviation capturing variability. Surface pressure is converted to hectopascals, with mean and standard deviation reflecting atmospheric stability. Wind speed is derived from 10-meter wind components (u10, v10), with mean and maximum values indicating prevailing and extreme conditions. Precipitation is converted to millimeters, with metrics including total precipitation, mean, standard deviation, number of rainy days (threshold: 0.1 mm), and intensity (total precipitation divided by rainy days). Specialized indices include the aridity index [15] (total precipitation divided by mean temperature plus 10) and the temperature-humidity index [16] (TH-index, combining temperature and humidity effects), capturing environmental interactions relevant to agriculture and human comfort. These features are aggregated into a pandas DataFrame with 34 rows (provinces) and 19 columns (features).

C. Data Preprocessing and Dimensionality Reduction

To address outliers and varying feature scales, data is scaled using scikit-learn's [12] RobustScaler, which employs the median and interquartile range for robustness against extreme values. PCA [7] is applied to the scaled data to reduce dimensionality, targeting 95% variance retention. The explained variance ratio for each principal component is computed, and feature loadings for the first two components (PC1, PC2) are analyzed to identify dominant climate drivers. This step ensures computational efficiency while preserving critical information for subsequent analyses.

D. Clustering and Classification

Unsupervised clustering is performed using K-means [17], [18] and hierarchical clustering (AgglomerativeClustering with Ward linkage [19]) for 3 to 7 clusters. The silhouette score [20], a measure of cluster cohesion and separation, evaluates clustering quality, with the highest score indicating the optimal configuration over the tested range. Supervised classification models, including K-nearest neighbors (KNN) [21], random forest [8], and gradient boosting [22], are trained on the scaled data using 5-fold cross-validation to assess predictive performance. Model hyperparameters are tuned to optimize accuracy, and cross-validation scores provide robust estimates of generalization.

E. Visualization

Thirteen visualizations are generated to interpret results, including a correlation heatmap, PCA explained variance plot, feature loadings plots, clustering comparison, PCA scatter plot, model performance comparison, distribution plots for key features (temperature, precipitation, humidity, wind), a per-class feature summary chart (mean ± standard deviation), a hierarchical clustering dendrogram, and a radar chart of climate class characteristics. These visualizations, saved in plots_dir, enhance the interpretability of complex patterns, facilitating communication with stakeholders.

IV. EXPERIMENTAL SETUP

A. Dataset

The dataset comprises meteorological data for 34 Chinese provinces, sourced from the ERA5 single-levels dataset [4]. Each province's data is stored in NetCDF format, containing variables such as temperature, dewpoint, pressure, wind components, and precipitation. The enhanced_climate_parameters function processes these variables to extract 19 features, resulting in a pandas DataFrame with 34 rows (provinces) and 19 columns (features). The dataset captures the diverse climatic conditions across China, from the humid subtropical south to the arid northwest.

B. Implementation Details

The pipeline is implemented in Python 3.10 within Google Colab, leveraging libraries such as xarray (for NetCDF processing), pandas (for data manipulation), scikit-learn [12] (for machine learning), and matplotlib/seaborn [13] (for visualization). Data preprocessing uses RobustScaler to standardize features, and PCA is configured to retain 95% variance. K-means clustering employs random initialization with 10 runs to ensure stability, while hierarchical clustering uses Ward linkage for minimizing variance. Classification models are trained with default hyperparameters, except for KNN (k=5) and random forest (100 trees), with 5-fold cross-validation ensuring robust performance estimates. Visualizations are generated at 300 DPI for publication quality, stored in plots_dir.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

V. RESULTS

The analysis yields comprehensive insights into climate patterns across Chinese provinces. Results are organized into numerical outputs, tables, and visualizations. Each subsection presents a specific result accompanied by a detailed explanation of its implications for climate classification and environmental applications.

A. Dataset Summary

The processed dataset comprises 34 provinces and 19 climate features (shape: 34×19), encapsulating the meteorological diversity of China. The dataset summary for the first five provinces (HeBei, Liaoning, Fujian, Shanxi, Jiangsu) is presented in three tables: Table I (temperature metrics), Table II (humidity and pressure), and Table III (wind and precipitation). These tables collectively highlight the variability in climate parameters, which is critical for distinguishing regional climate zones.

TABLE I SUMMARY OF CLIMATE FEATURES FOR FIRST FIVE PROVINCES (TEMPERATURE)

Province	Temp Mean (°C)	Temp Max (°C)	Temp Min (°C)	Temp Std	Temp Range
HeBei	14.50	42.69	−17.40	12.23	60.09
Liaoning	10.03	35.32	−29.91	13.80	65.22
Fujian	20.10	37.83	−1.85	7.32	39.68
Shanxi	10.37	36.29	−25.48	11.76	61.77
Jiangsu	17.00	36.93	−8.69	9.71	45.61

Table I details temperature metrics. Fujian exhibits a high mean temperature (20.10°C), indicative of a warm, subtropical climate, while Liaoning's wide temperature range (65.22°C) reflects continental extremes. Shanxi's lower mean temperature (10.37°C) and high standard deviation (11.76°C) suggest significant seasonal variability, typical of inland temperate regions.

TABLE II SUMMARY OF CLIMATE FEATURES FOR FIRST FIVE PROVINCES (HUMIDITY AND PRESSURE)

Province	Humidity Mean (%)	Humidity Std	Pressure Mean (hPa)	Pressure Std
HeBei	58.63	23.09	998.21	16.09
Liaoning	61.98	21.67	1000.81	11.03
Fujian	79.64	13.05	982.02	16.77
Shanxi	54.89	23.20	802.29	15.03
Jiangsu	73.43	18.72	1013.51	9.88

Table II presents humidity and pressure metrics. Fujian's high mean humidity (79.64%) and low standard deviation (13.05%) indicate consistently humid conditions characteristic of coastal subtropical zones. Shanxi's lower mean humidity (54.89%) and high variability (23.20%) reflect drier, more variable conditions due to its inland location. Shanxi's low mean pressure (802.29 hPa) is consistent with higher altitude, while Jiangsu's high mean pressure (1013.51 hPa) and low variability (9.88 hPa) suggest stable atmospheric conditions near sea level.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

TABLE III SUMMARY OF CLIMATE FEATURES FOR FIRST FIVE PROVINCES (WIND AND PRECIPITATION)

Province	Wind Mean	Wind Max	Precip Total (mm)	Precip Mean	Precip Std	Intensity	Aridity Idx
HeBei	2.06	8.86	5256.26	0.07	0.39	3.51	214.57
Liaoning	3.61	14.14	6055.78	0.08	0.46	3.87	302.39
Fujian	2.45	15.32	15734.78	0.10	0.78	3.82	522.16
Shanxi	2.84	9.06	1586.56	0.09	0.18	3.80	260.55
Jiangsu	2.81	8.95	11534.60	0.15	0.61	3.44	427.26

Table III covers wind and precipitation metrics. Fujian's high precipitation total (15,734.78 mm) and aridity index (522.16) underscore its wet, monsoon-driven climate, while Shanxi's low precipitation (1,586.56 mm) indicates arid conditions. Jiangsu's high precipitation mean (0.15 mm/day) and intensity (3.44 mm/day) suggest frequent, moderate rainfall. Wind metrics show Liaoning's high mean wind speed (3.61 m/s), reflecting exposure to coastal winds.

B. PCA Analysis

Principal component analysis reduces the 19-dimensional feature space to a lower-dimensional representation, with the first principal component explaining 47.6% and the second 19.6% of the total variance; five components are retained to capture approximately 95% of the variance. Table IV lists the five features with the largest contributions to PC1: pressure standard deviation (−0.476), precipitation standard deviation (0.335), TH-index (0.266), mean temperature (0.260), and mean wind speed (0.247). PC1 is therefore dominated by pressure variability and precipitation variability, reflecting the combined influence of topography and monsoon-driven rainfall regimes. Precipitation intensity, which loads only weakly on PC1, instead contributes to PC2 (0.188), an axis governed chiefly by thermal extremes and pressure (maximum temperature 0.513, pressure standard deviation 0.479, pressure mean 0.417).

TABLE IV TOP FIVE FEATURES FOR PC1 LOADINGS

Feature	PC1 Loading
Pressure Std	−0.476
Precipitation Std	0.335
TH Index	0.266
Temperature Mean	0.260
Wind Mean	0.247

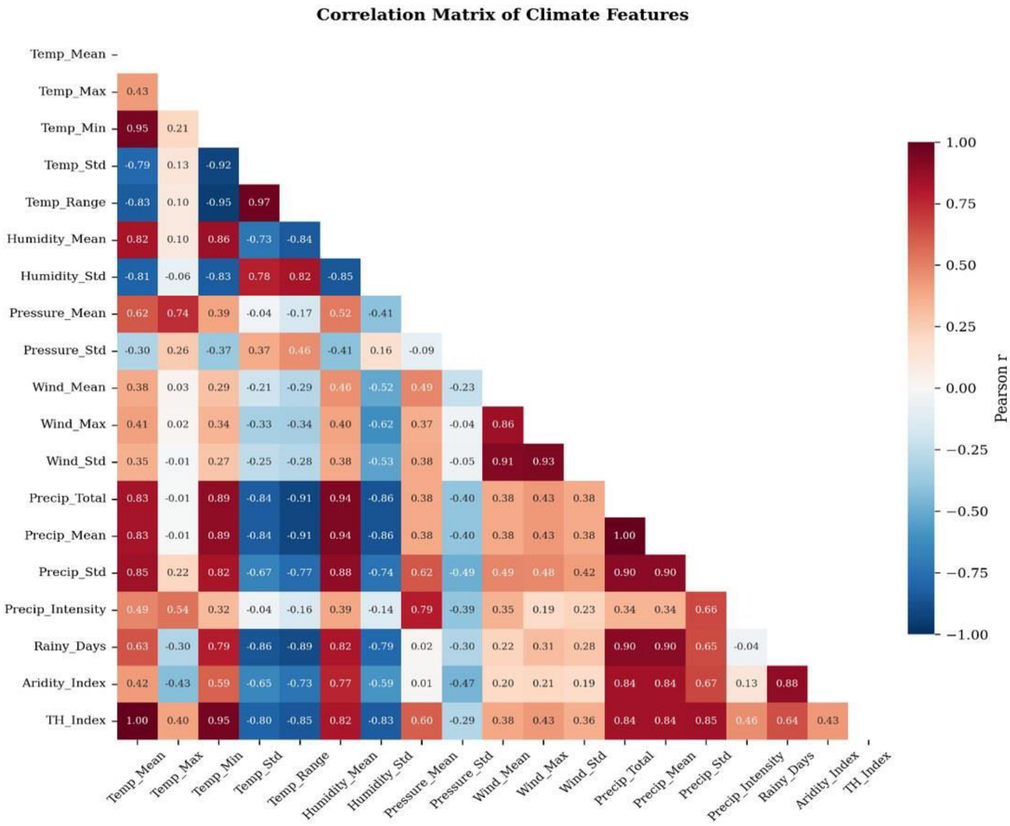


Fig. 1. Correlation Matrix of Climate Features

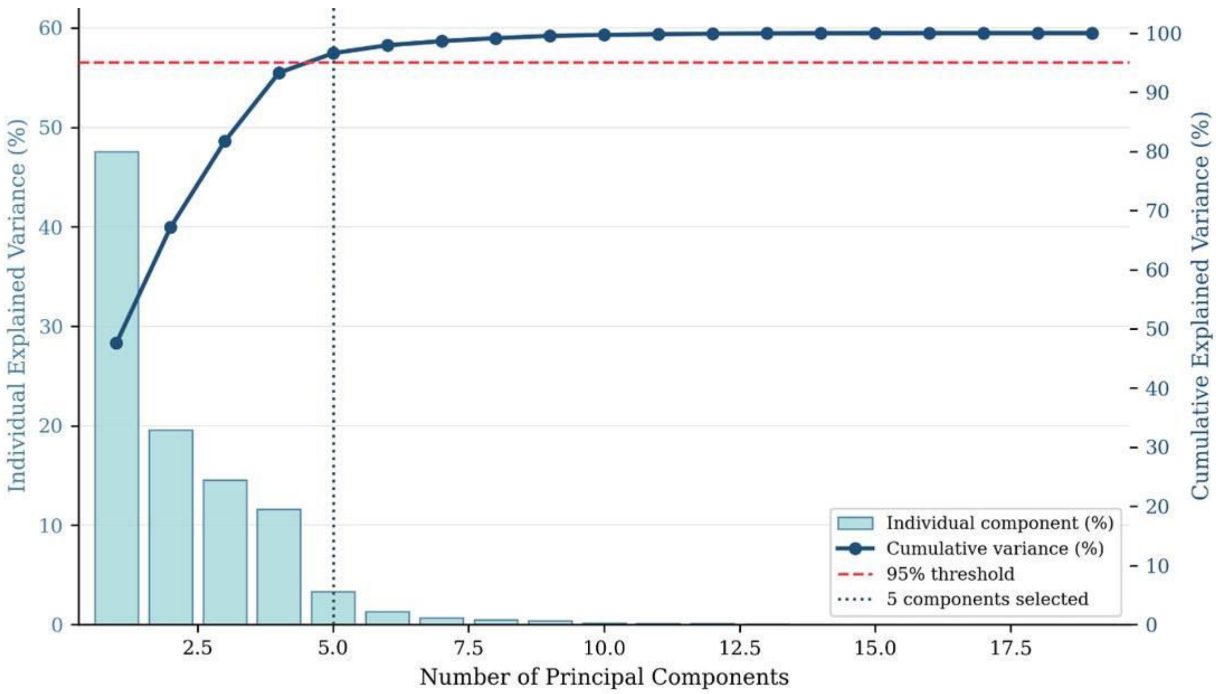


Fig. 2. PCA Explained Variance Analysis

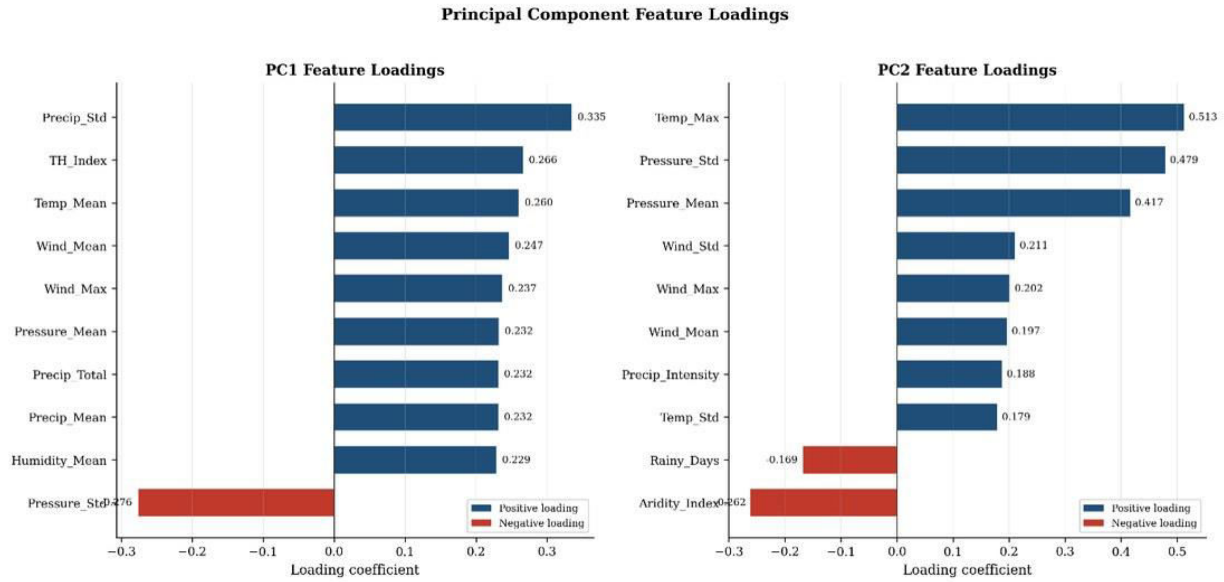


Fig. 3. Principal Component Feature Loadings (PC1 and PC2)

C. Clustering Results

Clustering analysis evaluates K-means and hierarchical clustering for 3 to 7 clusters, with silhouette scores reported in Table V. K-means with seven clusters attains the highest silhouette score of 0.350, while hierarchical clustering peaks at 0.319 for three clusters. The K-means silhouette increases monotonically across the tested range (0.255 at k = 3 to 0.350 at k = 7), indicating that the data do not exhibit a single sharply preferred cluster count; the seven-cluster solution is selected as the best-scoring configuration over this range. A silhouette of this magnitude denotes moderate, partially overlapping separation rather than strongly isolated clusters, consistent with the continuous nature of climatic gradients across China.

TABLE V CLUSTERING PERFORMANCE (SILHOUETTE SCORES)

Number of Clusters	K-Means Silhouette	Hierarchical Silhouette
3	0.255	0.319
4	0.316	0.307
5	0.319	0.307
6	0.336	0.269
7	0.350	0.323

The resulting classes broadly track geography: a hot, very wet tropical-coastal group in the far south (e.g., Hainan, Taiwan, Macau, Hong Kong), a large humid-subtropical group across the central and southern provinces, a northern temperate group, a cold and dry interior group in the northwest, and high-altitude and arid outliers such as the Tibetan Plateau (Xizang, Qinghai) and Xinjiang. The hierarchical clustering dendrogram (Fig. 7) illustrates these similarities, with branch lengths indicating feature disparities.

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406002
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

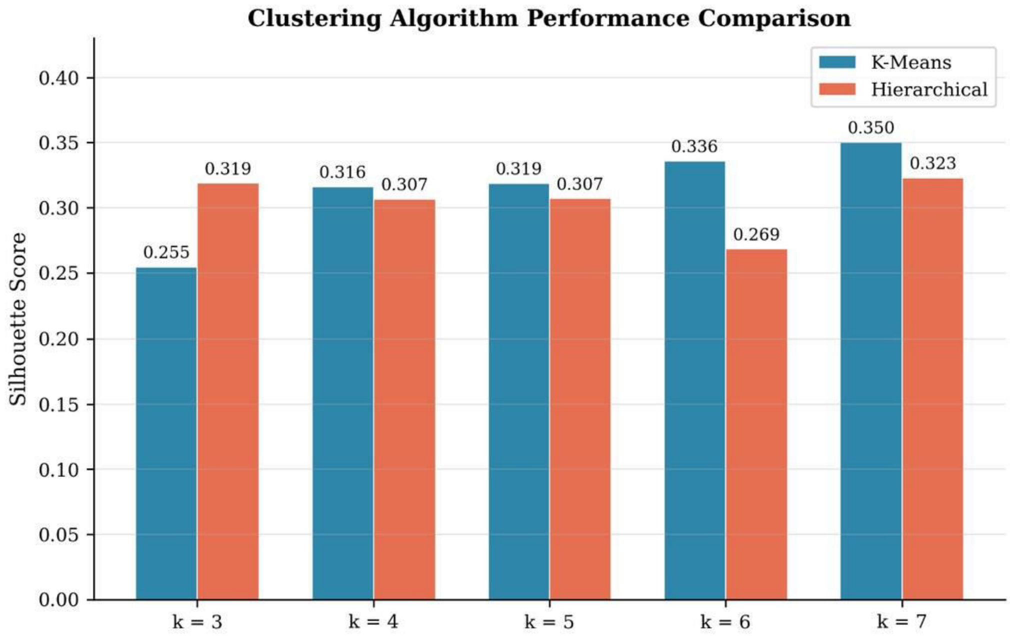
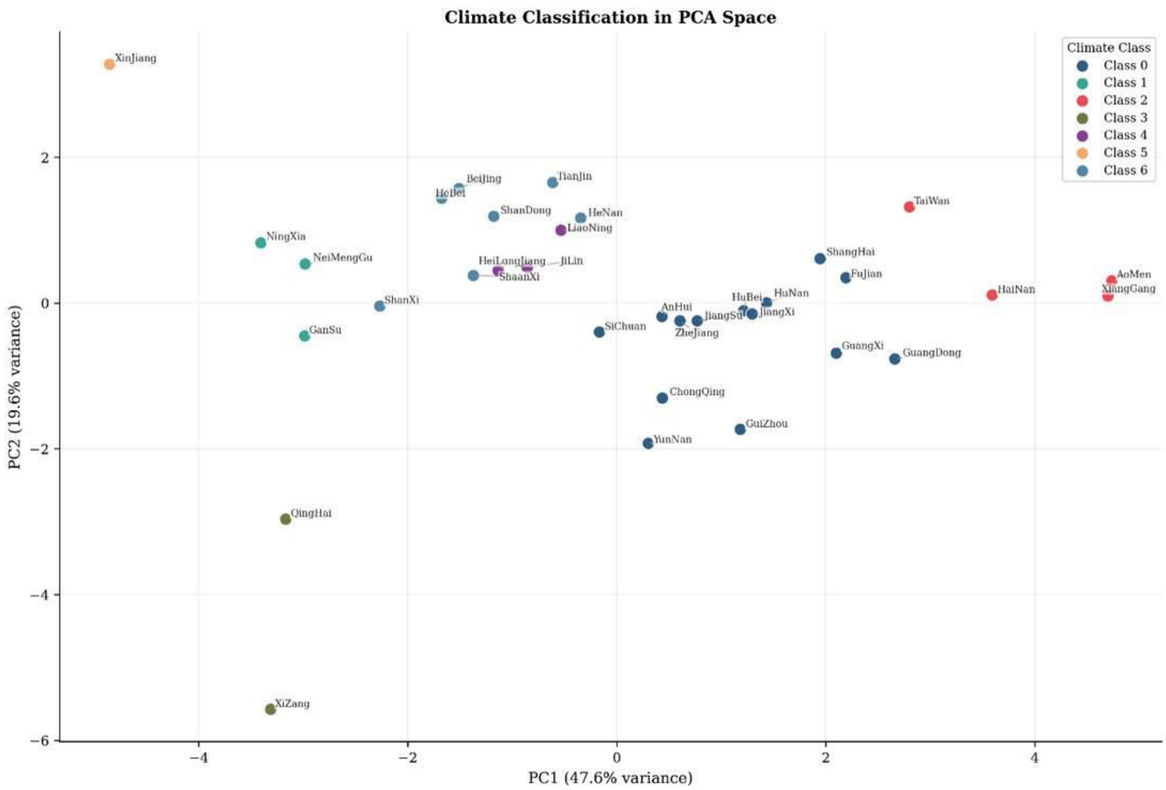


Fig. 4. Clustering Algorithm Performance Comparison



Classification Model Performance (5-Fold CV)

Fig. 5. Climate Classification in PCA Space

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406002
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

D. Classification Performance

Classification models are trained to predict climate-class membership, with 5-fold cross-validation performance shown in Table VI. Gradient boosting achieves the highest mean accuracy (0.771), followed by KNN (0.762) and random forest (0.738). Random forest shows the most consistent performance across folds, whereas gradient boosting exhibits the widest fold-to-fold variability. This variability is expected: partitioning only 34 provinces into seven classes leaves several classes with very few members, so each cross-validation fold sees a markedly different class balance, which destabilizes the more flexible ensemble learners.

TABLE VI CLASSIFICATION MODEL PERFORMANCE

Model	Mean CV Score	Std CV Score
Gradient Boosting	0.771	0.146
KNN	0.762	0.078
Random Forest	0.738	0.052

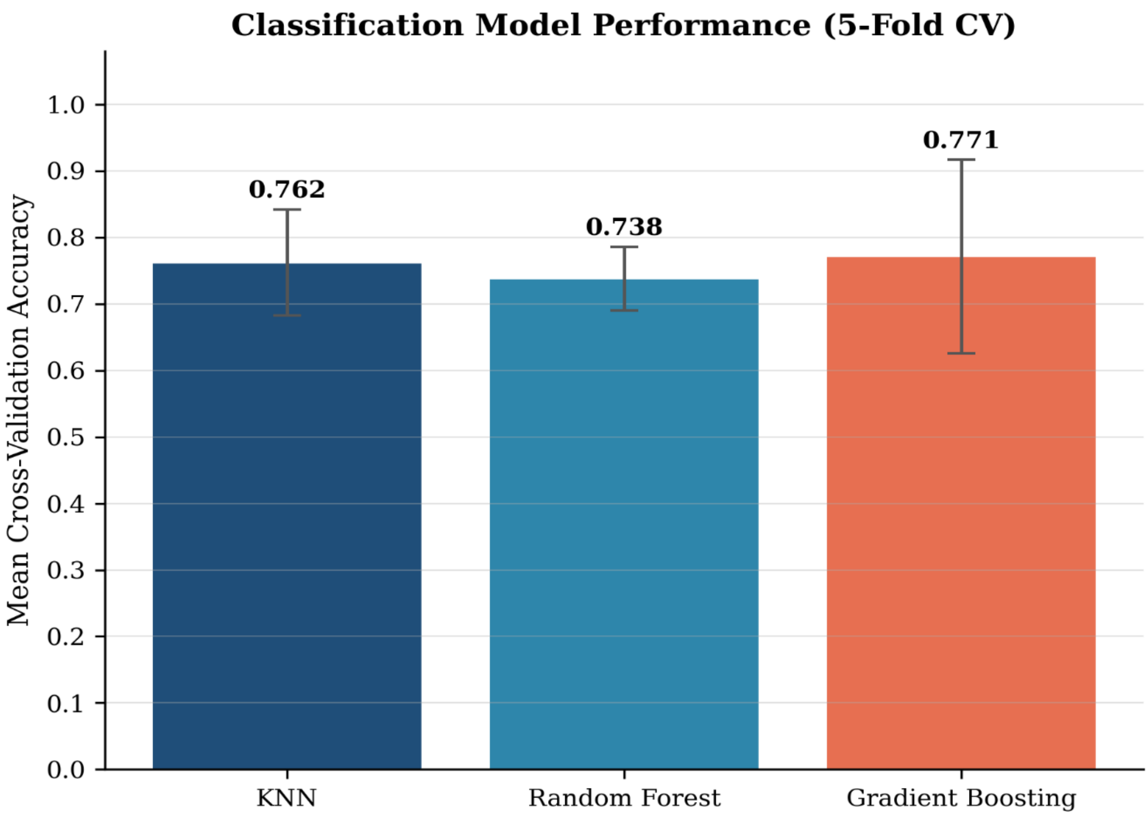


Fig. 6. Classification Model Performance (5-Fold CV)

E. Visualizations

Thirteen visualizations provide a comprehensive interpretation of the analysis. The correlation heatmap (Fig. 1) shows that total precipitation correlates almost perfectly with mean precipitation (r = 1.00) and strongly with precipitation standard deviation (r = 0.90) and the aridity index (r = 0.84), whereas precipitation intensity correlates only moderately with totals (r ≈ 0.34). Mean temperature is positively associated with mean pressure (r = 0.62); the strongest negative correlations involve temperature range with mean and minimum temperature (r = −0.83 and −0.95) and with precipitation totals (r = −0.91), capturing the contrast between mild, wet, low-seasonality climates and cold, dry, high-seasonality ones. The PCA explained-variance plot (Fig. 2) confirms that five components account for approximately

© 2026 IJIRCCE | Volume 14, Issue 6, June 2026 DOI: 10.15680/IJIRCCE.2026.1406002
www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|
International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)
(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

95% of the variance, and the feature-loadings plot (Fig. 3) reinforces pressure and precipitation variability as the dominant PC1 drivers, with maximum temperature and pressure terms loading most heavily on PC2.

The clustering-comparison plot (Fig. 4) summarizes silhouette scores across k = 3–7 and supports the selection of seven clusters for K-means as the best-scoring configuration. The PCA scatter plot (Fig. 5) depicts the provinces in PC1–PC2 space, with the seven climate classes occupying distinct regions. The hierarchical clustering dendrogram (Fig. 7) visualizes province similarities. Distribution plots for mean temperature (Fig. 8), total precipitation (Fig. 9), mean humidity (Fig. 10), and mean wind speed (Fig. 11) illustrate how the classes differ across key variables. The feature-summary chart (Fig. 12) reports the mean and standard deviation of each key feature by class, and the radar chart (Fig. 13) compares normalized feature values across the seven climate classes.

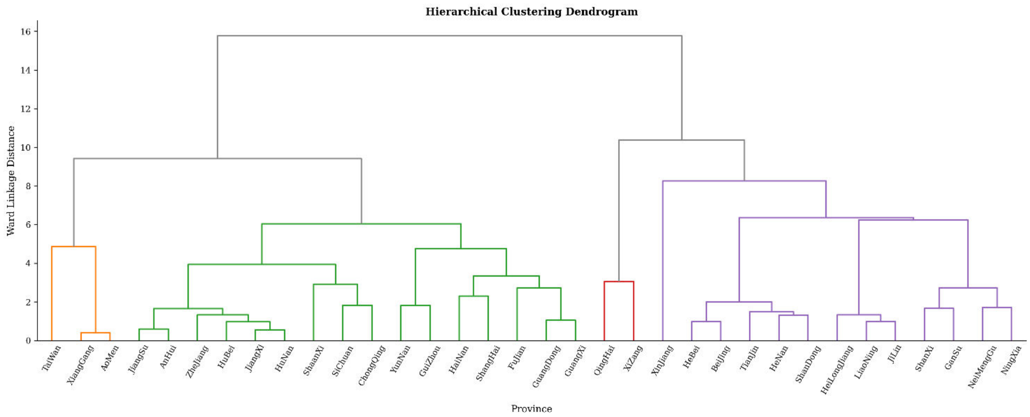


Fig. 7. Hierarchical Clustering Dendrogram

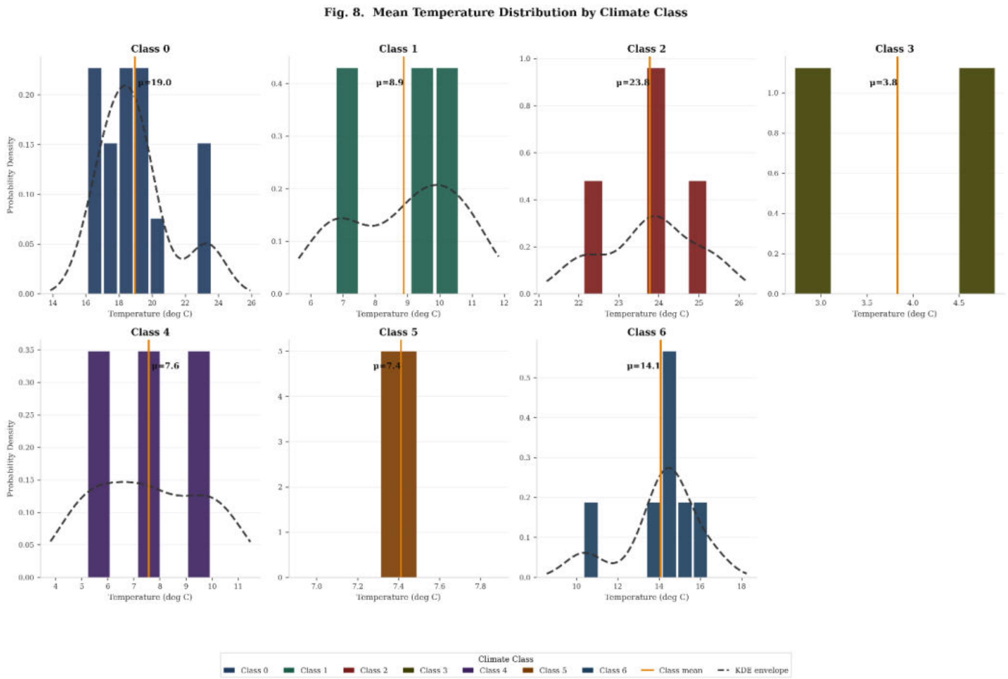


Fig. 8. Mean Temperature Distribution by Climate Class

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

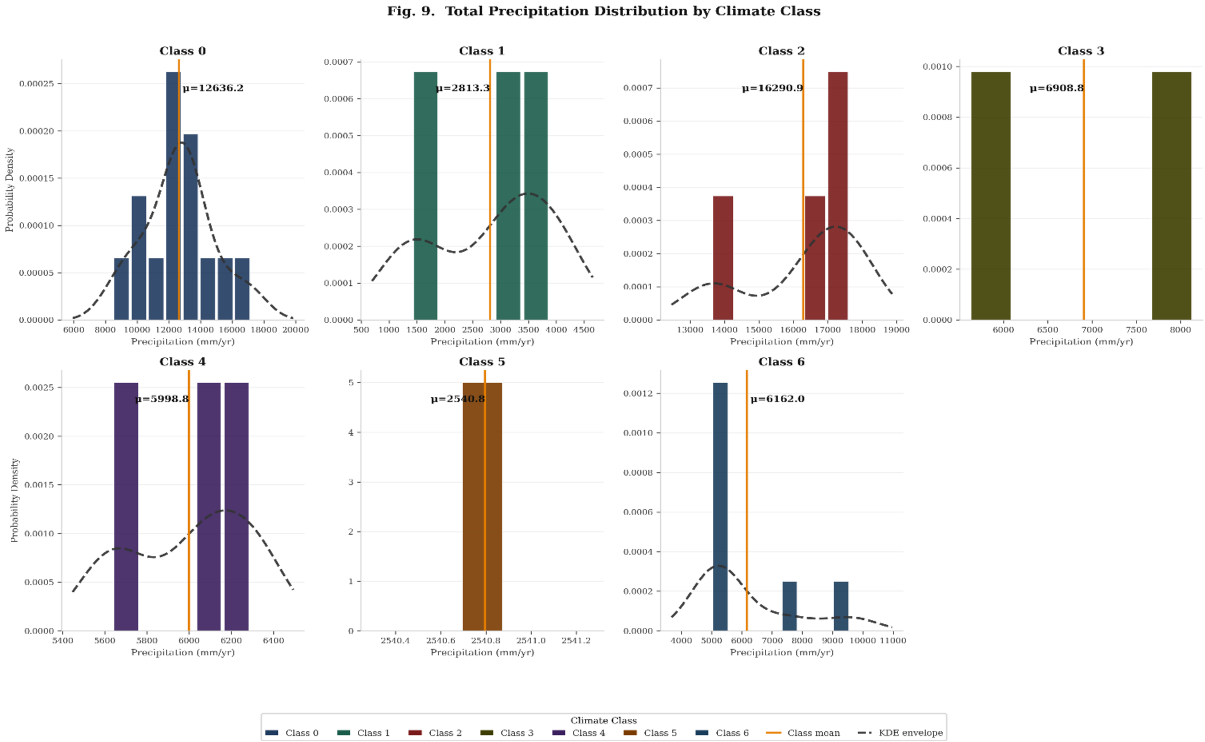


Fig. 9. Total Precipitation Distribution by Climate Class

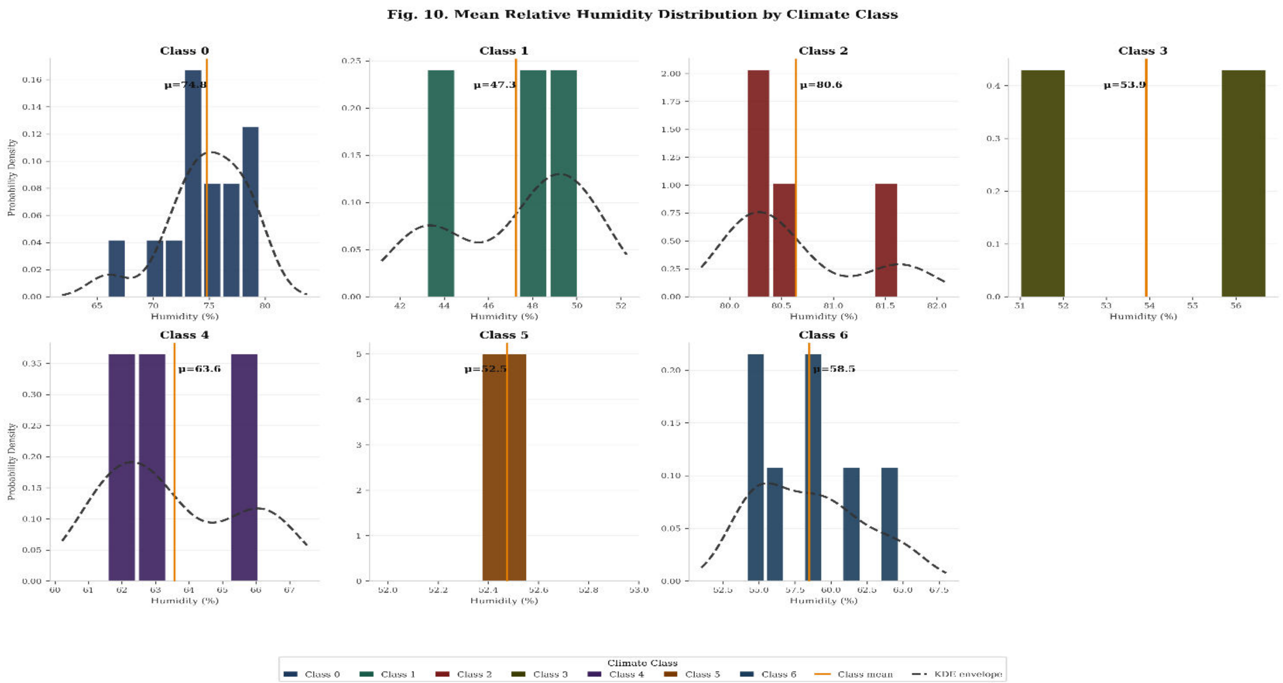


Fig. 10. Mean Relative Humidity Distribution by Climate Class

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

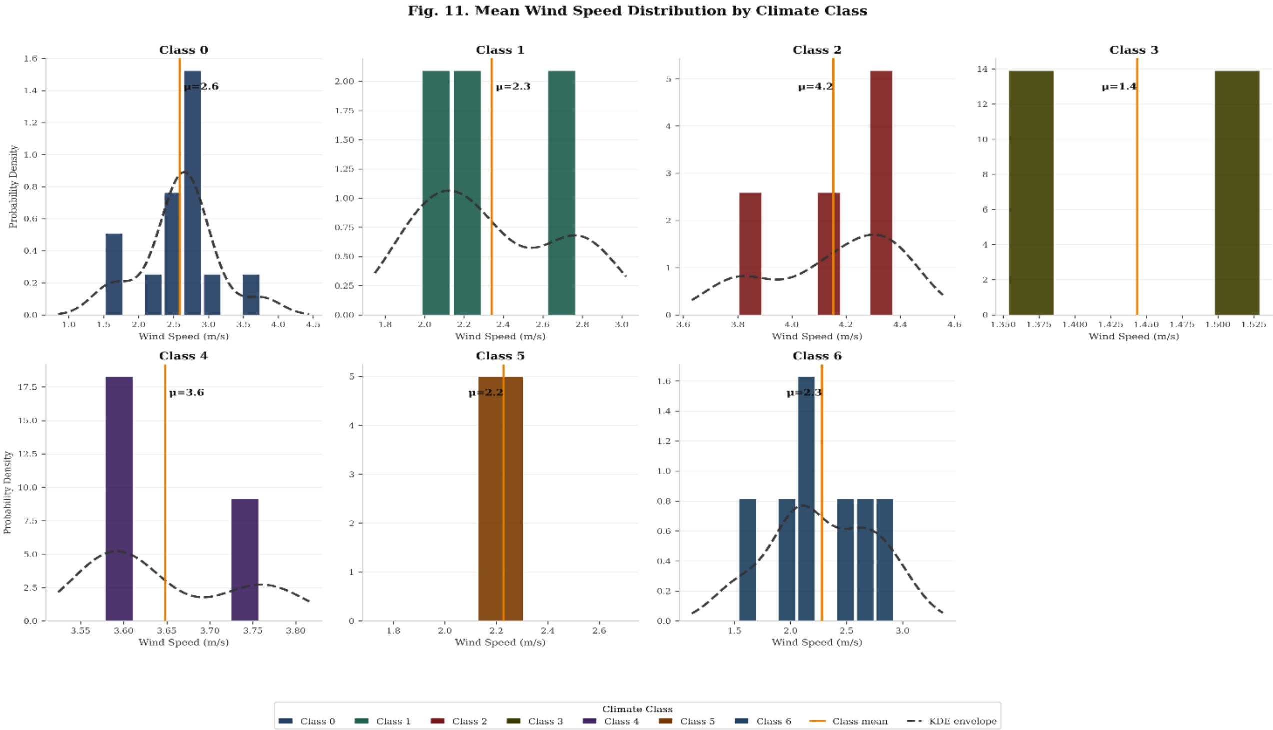


Fig. 11. Mean Wind Speed Distribution by Climate Class

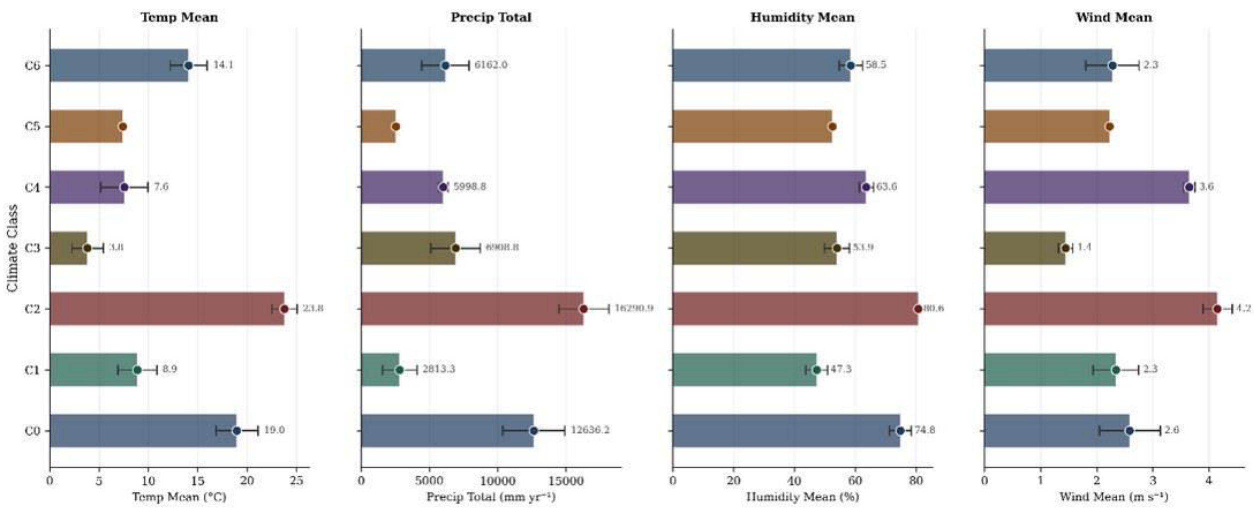


Fig. 12. Climate Feature Summary by Class (Mean ± Std)

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

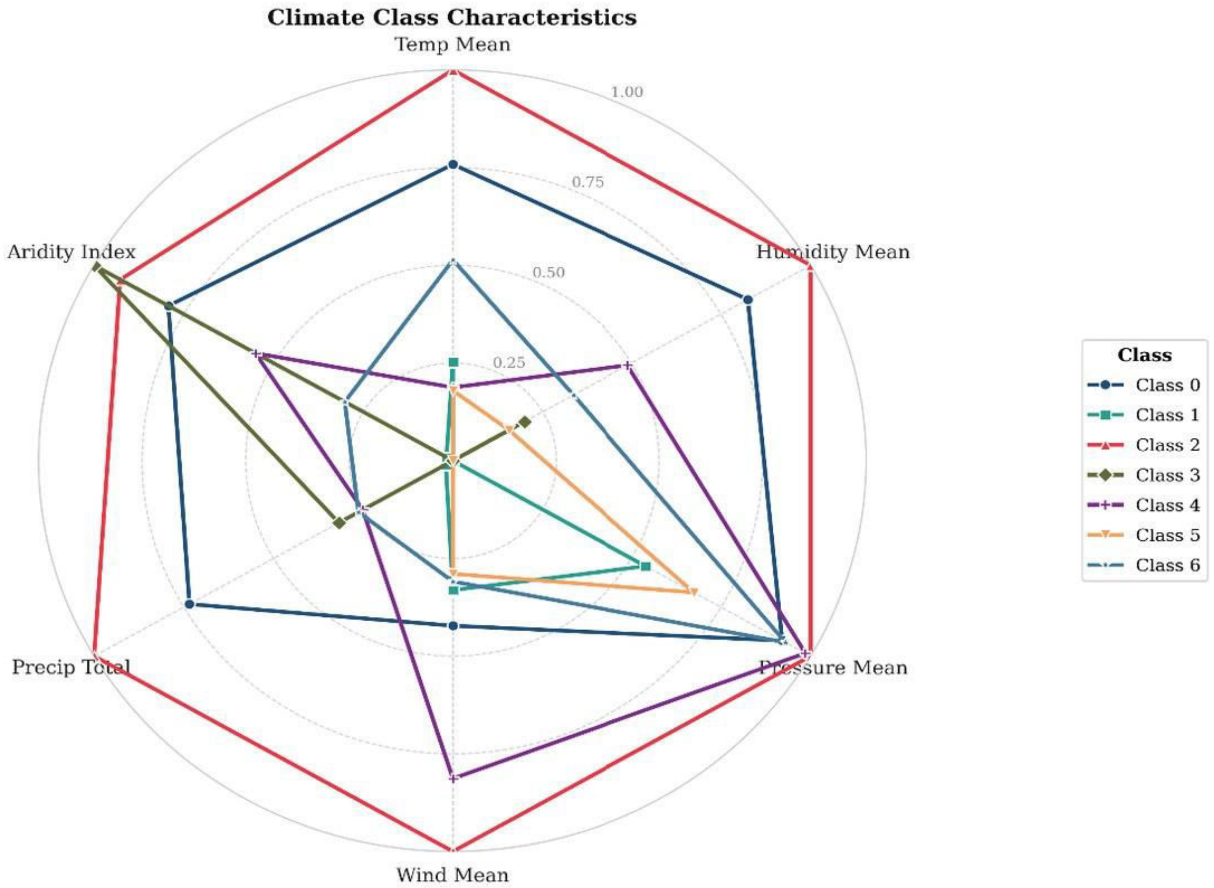


Fig. 13. Climate Class Characteristics (Radar Chart)

VI. DISCUSSION

The results demonstrate the efficacy of the proposed machine learning pipeline in characterizing climate across Chinese provinces, offering several key insights. The dataset's 19 features, derived through advanced engineering, capture the multifaceted nature of climate, enabling differentiation of the provinces into seven classes via K-means clustering (silhouette score: 0.350). The moderate silhouette score indicates discernible but partially overlapping clusters, reflecting the smooth climatic gradients of the region, for example, the separation of hot, humid coastal provinces (e.g., Fujian, Hainan) from cold, arid interior regions (e.g., the Tibetan Plateau, Xinjiang). This data-driven grouping complements traditional systems like Köppen-Geiger [1], which may group diverse climates under broad categories, and is consistent with recent clustering studies of Chinese climate that report comparable advantages over threshold-based schemes [2], [3].

The PCA analysis reveals that pressure variability and precipitation variability are the dominant drivers along the leading component, reflecting the influence of topographic variations and monsoon dynamics, while thermal extremes and pressure govern the second component. The classification accuracy of gradient boosting (0.771) and KNN (0.762) supports reliable, though not exceptional, prediction of climate classes for new data. The visualizations, particularly the radar chart and PCA scatter plot, provide intuitive representations of climate profiles, facilitating communication with stakeholders in environmental management.

However, the study acknowledges limitations. With seven classes drawn from only 34 provinces, several classes contain very few members (some are effectively singletons, such as the high-altitude Tibetan Plateau and the arid northwest), which destabilizes cross-validation estimates and inflates the variance of the ensemble classifiers.

Moreover, the K-means silhouette increases monotonically across the tested range, so the seven-class solution is the best-scoring configuration over k = 3–7 rather than a strongly separated natural optimum; a larger, finer-grained sample would be needed to establish a stable cluster count. The ERA5 dataset, while high-resolution, may introduce minor uncertainties due to reanalysis interpolation, particularly in mountainous regions. Future work could address these limitations by incorporating temporal trends to capture climate change impacts, additional variables (e.g., soil moisture, vegetation indices), and larger datasets to enhance granularity. Integrating deep learning models, such as neural networks, could further improve classification accuracy, particularly for complex non-linear patterns.

The framework's applications are extensive. In agriculture, the climate classes guide crop selection and irrigation strategies tailored to regional precipitation and temperature profiles. In urban planning, understanding humidity and wind patterns informs infrastructure design for climate resilience. For policymakers, the visualizations provide evidence-based insights for climate adaptation strategies, such as flood preparedness in high-precipitation regions like Fujian or drought mitigation in arid areas like Shanxi. This study's scalable approach positions it as a valuable tool for regional climate analysis, with potential for adaptation to other countries globally.

VII. CONCLUSION

This study develops an enhanced climate classification framework for 34 Chinese provinces using the ERA5 dataset, leveraging a machine learning pipeline that integrates feature engineering, PCA, clustering, and classification. The identification of seven distinct climate classes via K-means clustering, with a silhouette score of 0.350, captures the broad regional climate variability of China while reflecting the continuous nature of its climatic gradients. Gradient boosting attains the highest classification accuracy (0.771), with KNN close behind (0.762), validating the predictive value of the derived features. Thirteen visualizations, including correlation heatmaps, PCA scatter plots, and radar charts, provide actionable insights into feature interactions and climate class characteristics, with pressure variability and precipitation variability emerging as dominant factors. The framework offers a scalable, data-driven approach for regional climate analysis, with applications in environmental planning, agriculture, and policy-making. Future work could enhance the framework by incorporating temporal dynamics, additional environmental variables, and deep learning techniques to further refine classifications and address emerging climate challenges.

REFERENCES

[1] W. Köppen, “Das geographische System der Klimate,” in Handbuch der Klimatologie, vol. 1, Berlin, Germany: Gebrüder Borntraeger, 1936.
[2] J. Shi and L. Yang, “A climate classification of China through k-nearest-neighbor and sparse subspace representation,” Journal of Climate, vol. 33, no. 1, pp. 243–262, 2020.
[3] H. Duan, Q. Li, L. He, J. Zhang, H. An, R. Ali, and M. Vazifedoust, “Climate classification for major cities in China using cluster analysis,” Atmosphere, vol. 15, no. 7, art. no. 741, 2024.
[4] Copernicus Climate Change Service, “ERA5 reanalysis single-levels dataset,” Copernicus Climate Data Store, 2025. [Online]. Available: https://cds.climate.copernicus.eu/datasets/reanalysis-era5-single-levels
[5] M. C. Peel, B. L. Finlayson, and T. A. McMahon, “Updated world map of the Köppen-Geiger climate classification,” Hydrology and Earth System Sciences, vol. 11, pp. 1633–1644, 2007.
[6] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, “Present and future Köppen-Geiger climate classification maps at 1-km resolution,” Scientific Data, vol. 5, art. no. 180214, 2018.
[7] I. T. Jolliffe, Principal Component Analysis, 2nd ed. New York, NY, USA: Springer-Verlag, 2002.
[8] L. Breiman, “Random forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[9] H. Hersbach, B. Bell, P. Berrisford, et al., “The ERA5 global reanalysis,” Quarterly Journal of the Royal Meteorological Society, vol. 146, no. 730, pp. 1999–2049, 2020.
[10] D. P. Dee, S. M. Uppala, A. J. Simmons, et al., “The ERA-Interim reanalysis: configuration and performance of the data assimilation system,” Quarterly Journal of the Royal Meteorological Society, vol. 137, no. 656, pp. 553–597, 2011.
[11] C. R. Harris, K. J. Millman, S. J. van der Walt, et al., “Array programming with NumPy,” Nature, vol. 585, no. 7825, pp. 357–362, 2020.
[12] F. Pedregosa, G. Varoquaux, A. Gramfort, et al., “Scikit-learn: machine learning in Python,” Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011.

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 9.927| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

[13] J. D. Hunter, "Matplotlib: a 2D graphics environment," Computing in Science & Engineering, vol. 9, no. 3, pp. 90–95, 2007.
[14] O. A. Alduchov and R. E. Eskridge, "Improved Magnus form approximation of saturation vapor pressure," Journal of Applied Meteorology, vol. 35, no. 4, pp. 601–609, 1996.
[15] E. de Martonne, "L'indice d'aridité," Bulletin de l'Association de Géographes Français, vol. 9, no. 3, pp. 3–5, 1926.
[16] E. C. Thom, "The discomfort index," Weatherwise, vol. 12, no. 2, pp. 57–61, 1959.
[17] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in Proc. 5th Berkeley Symp. on Mathematical Statistics and Probability, vol. 1. Berkeley, CA, USA: Univ. of California Press, 1967, pp. 281–297.
[18] S. P. Lloyd, "Least squares quantization in PCM," IEEE Transactions on Information Theory, vol. 28, no. 2, pp. 129–137, 1982.
[19] J. H. Ward, "Hierarchical grouping to optimize an objective function," Journal of the American Statistical Association, vol. 58, no. 301, pp. 236–244, 1963.
[20] P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," Journal of Computational and Applied Mathematics, vol. 20, pp. 53–65, 1987.
[21] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," IEEE Transactions on Information Theory, vol. 13, no. 1, pp. 21–27, 1967.
[22] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189–1232, 2001.