INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Comparative Analysis of Traditional Time Series and Machine Learning Models for Multi-Year Temperature Forecasting

N Dhwani[1,2], P Pranay[1]*, N Sharanya[3]

1 Chaitanya (Deemed to be University), Warangal, Telangana, India
2 Little Flower Degree College, Uppal, Hyderabad, India
3 Vasundara Women's Degree College, ECIL, Hyderabad, India

OPEN ACCESS

Received: 17/07/2025

Accepted: 16/09/2025

Published: 29/10/2025

Citation: Dhwani N, Pranay P, Sharanya N (2025) Comparative Analysis of Traditional Time Series and Machine Learning Models for Multi-Year Temperature Forecasting. Indian Journal of Science and Technology 18(39): 3145-3158. https ://doi.org/10.17485/IJST/v18i39.1263

* **Corresponding author**.

pettempranay@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Forecasting temperatures accurately is important for sectors that are sensitive to temperatures (agriculture, disaster management, public health, environment). **Objective:** Given an increase in variability attributed to climate change, the aim of this study is to comparatively analyse the traditional time series analysis methods (ARIMA and SARIMA) against the current popular and improved machine learning and deep learning methods (Multiple Linear Regression, Decision Tree, Random Forest, Artificial Neural Network, Long Short-Term Memory). **Method:** The analysis draws on daily temperature observations from 2017 to 2024 obtained from the India Meteorological Department (IMD), Begumpet, Hyderabad. Each of the models are evaluated using the Root Mean Square Error (RMSE) metric. **Findings:** Among the traditional models, SARIMA outperformed ARIMA with RMSE scores ranging from 0.69 (2019) to 2.90 (2024); ARIMA had RMSE scores of 1.92 (2019) and ranged to as high as 6.33 (2023). However, both traditional models were inferior to all of the machine learning models. The MLR model was the most accurate with the lowest RMSE score of 0.2921 (2020) and at the most 0.7239 (2023). In terms of R^2 values, the best performing model is a Multiple Linear Regression (MLR), which repeatedly achieved the highest and most stable values of 0.9749 to 0.9918 in all years. Random Forest (RF) was second, RMSE scores ranging from 0.4710 to 0.8623. ANN and LSTM showed variability in performance, LSTM with peaked RMSE at 2.4887 (2023) and ANN peaked at 3.1496 (2023). The analysis shows that non-complex machine learning models (MLR and RF) provided improved accuracy and robustness versus the traditional ARIMA based models, especially when temperature variability is high. **Novelty:** The paper is innovative as it presents a year-wise comparison of ARIMA, SARIMA and ML models in long-term temperature forecasting, demonstrating that basic ML models are better than traditional models using real climate data.

Keywords: Temperature Forecasting; RMSE; Decision Tree; Random Forest; Hyderabad

1 Introduction

Accurate temperature forecasting is of utmost importance to many industries, including agriculture, public health, disaster response, and environmental planning. Climate change continues to alter weather patterns and add to the variability and complexity faced in outstandingly important and life-dependent industries. The recent years have placed considerable focus on determining models that can seasonally adapt to nonlinear and abrupt variability in temperature data. Thus, models must be adaptable to explain the increasingly random and unexpected temperature process. Temperature data consisting of observations from the years of 2017 to 2024 were collected from the Meteorology Department of Begumpet Hyderabad, a reliable source for official climate data. The data set includes daily temperature readings, providing us with a dataset to evaluate forecasting methods in the real-world. The objective of this research is to compare the two traditional time series methods identified as (Autoregressive Integrated Moving Average (ARIMA)and Seasonal ARIMA (SARIMA)) and further develop it with recent physical machine learning methods in the form of multiple linear regression (MLR), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN), and Long Short-Term Memory (LSTM). The methods will be evaluated using the Root Mean Squared Error (RMSE) with multiple years to understand how robust and reliable different forecasting methods are at forecasting temperature. These results help to inform meteorologists, data scientists, and decision makers about what techniques will be appropriate for use in forecasting climate related decisions. Using data obtained from a single consistent location and evaluating models across several years, allows this study to accurately depict challenges faced in realistic forecasting, but also offered points of practical magnitude to think about for model selection related to urban temperature forecasting. [(1)] Demonstrated that in predictive tasks, machine learning models performed up to 25% more accurately than conventional techniques. Convolution Neural Network (CNN) -LSTM performed better than ARIMA in temperature forecasting, according to [(2)], with a lower RMSE (1.38°C vs. 1.56°C). Random Forest demonstrated superior accuracy with weather-influenced time series data, outperforming ARIMA in predicting H5N1 outbreaks with a significantly lower MSE. An enhanced Recurrent Neural Network (RNN) model outperformed ARIMA in COVID-19 case forecasting with 91% accuracy and a lower RMSE (2,016 vs. 4,391), demonstrating the promise of deep learning in epidemic modeling [(3)].

1.1 Literature Review

[(4)] Compared five machine learning classifiers-Random Forest, XGBoost, C5.0, Conditional Inference Tree, and Logistic Model Tree-for glaucoma diagnosis on Spectralis OCT data; of these methods, the Random Forest exhibited the best accuracy of 93.7% followed by XGBoost with 92.1%, much better than the logistic regression method. Macular ganglion cell layer and RNFL thickness were identified as the most predictive features for early glaucoma detection. The investigation demonstrated by [(5)] that REACH as a Random forest-based regression chain model, surpassed ARIMA and other conventional models when predicting influenza-like illness (ILI) cases. The REACH model demonstrated superior prediction accuracy through New Zealand's three flu seasons by producing a Mean Absolute Error (MAE) of 6.01, which was better than ARIMA's MAE of 7.13 during 1 to 7-day ahead forecasts. The prediction system demonstrated enhanced flexibility when dealing with both time-based and region-based changes in ILI patterns. [(3)] Discovered the RNNCON Res model, which was able to reach 91% accuracy for the forecasting of COVID-19/or Omicron cases twenty days ahead of the first confirmed case in France, presenting a significant improvement over conventional models. The RNNCON Res model achieved a 0.904 precision, 0.917 sensitivity, with an RMSE of 2,016, whilst ARIMA had an RMSE of 4,391. Therefore, the results demonstrate that the predictive performance of their model clearly outperformed classical time series methods. [(6)] Showed that machine learning models such as GBM were slightly better at predicting outcomes of heart failure than classic logistic regression models. GBM had C-statistics of 0.727 versus 0.724 for mortality and 0.745 vs. 0.707 for hospitalization. Both GBM and logistic regression were close, but showed marginal improvement over traditional classical models, putting practical use for advanced techniques into question. [(7)] Stated that their CNN-based deep learning method (VGG-16 + Random Forest classifier) outperformed classic machine-learning methods for facial emotion recognition featuring ageing adults, with at least 8% better accuracy in many different datasets (FACES, Lifespan, CIFE, and FER2013), and that deep architectures are particularly better at contending with age-related facial variability. The research findings of [(4)] showed that machine learning models, particularly Random Forest, produce better glaucoma detection results than traditional classifiers when using Spectralis OCT features. [(8)] stated that machine learning models showed only small improvements over logistic regression for in-hospital mortality prediction after cardiac surgery, with AUCs of 0.797 vs 0.791 (temporal) and 0.732 vs 0.713 (spatial), respectively; and concluded that ML's clinical benefits are modest and should be treated cautiously. [(9)] Concluded that when comparing machine learning and traditional statistical modelling in healthcare analytics, machine learning models performed approximately 12 % better than traditional models for selected tasks, but statistical modelling retains more interpretability and reliability, particularly on smaller datasets. [(7)] Concluded that their deep learning models based on the CNN model (VGG-16 +RF classifier) had at least 8% better accuracy than traditional methods for datasets on ageing adults; their accuracy varied between 87%-95% on FACES, Lifespan, CIFE, and FER2013, compared to classical algorithms that ranged from 79%-87%.[10] Compared machine learning (ML) to traditional statistical models for predicting undiagnosed diabetes using 32,827 participants from the KNHANES survey (2014–2020). The ML models with the highest performance achieved externally validated AUCs of 0.788 (versus 0.740) using four features, 0.802 (versus 0.759) using seven features, and 0.819 (versus 0.765) using eleven features, indicating the capacity of today's ML to outperform traditional models in all scenarios.[1] Compared traditional statistical models (such as linear regression) to more advanced machine learning methodologies (specifically: Random Forest, Decision Trees, and LSTM), which were just using historical stock price data to predict the stock price today, and concluded that ML models in his study performed better than traditional methods (using accuracy and reliability!). ML model accuracy and reliability were around 20-25% better than the traditional model.[1] Raised important issues regarding the interpretability of models and data privacy. The highest overall accuracy of neural network (CNN) for classifying ultrasound breast lesions was 91% and was followed by Random Forest (90%) and AutoML Vision (86%). AutoML achieved the best AUCPR score (0.95), highlighting the clinical relevance[11].[12] Showed that machine learning models yielded better demand forecasts than traditional methods. For example, Random Forest achieved an RMSE of 24.88 while ARIMA produced an RMSE of 31.71 and exponential smoothing had an RMSE of 34.95, thus showing that ML models provided superior accuracy on industrial datasets.[13] Demonstrated that machine learning algorithms provided better discrimination than traditional statistical models for predicting heart failure readmissions, and mortality. In the 21 time-point comparisons that were made, ML models had higher c-indices than traditional models in 16 cases. In the only externally validated study, the ML method reached a c-index of 0.913 while the traditional methods produced c-indices of 0.835 (pubmed.ncbi.nlm.nih.gov).[14] Concluded that machine learning and deep learning models provided greater accuracy than traditional statistical methods when forecasting the multivariate and nonlinear time series data. However, they stressed that a great deal more research needs to be conducted to understand the conditions under which machine learning methods provide better accuracy consistently on a variety of real-world datasets (scholar.smu.edu).

1.2 Research Gap

Past studies have used time series models (like ARIMA/SARIMA) and machine learning for temperature prediction separately. There is limited research on year-by-year comparisons over longer periods using actual weather data. Most studies focus on short-term predictions, multiple inputs, or how well one model works. They don't typically compare traditional and machine learning models side-by-side using the same data. Also, deep learning models, like LSTM and ANN, can act differently over time. This hasn't been looked at enough, especially in places like Hyderabad. Because of this, there is a need for studies that compare both types of models over many years to see how accurate and consistent they are.

1.3 Objectives

- To Analyze accuracy of traditional time series models vs machine learning for temperature forecasting
- To Compare models on daily temperature data (2017 to 2024) population from IMD Hyderabad
- To Determine or measure forecasting accuracy using Root Mean Squared Error (RMSE)
- To Identify the most precise and reliable temperature forecast model

2 Methodology

Data from the India Meteorological Department (IMD) at Begumpet Hyderabad Telangana regarding temperature forecasting for Hyderabad City between January 1, 2017, and December 31, 2024, consists of eight full years. The dataset contains daily measurements of essential weather parameters which consist of daily highs and lows temperature dew point humidity and rainfall and wind parameters and solar radiation. The study period was chosen because it shows both recent weather patterns as well as year-to-year variations and includes times of extreme weather conditions. https://docs.google.com/spreadsheets/d/1Yk86reo4DFqACmNWx5mGY0aMnXT1gdBX/edit?usp=drive_link&ouid=112858300260286252116&rtpof=true&sd=true. This paper offers an empirical evaluation of conventional time series models (ARIMA and SARIMA) against machine learning models for temperature time series forecasting, while using real-world data spanning the period of 2017 to 2024. The paper intends to distinguish performance in terms of accuracy metrics in order to define the most accurate and robust model for temperature prediction. An analysis of the results year by year is provided to see which model best learns and adapts to seasonality, variability, and climate change.

2.1 Structure of the paper

The flowchart depicts the systematic procedure used in this research, which starts with the initial collection of data and its preprocessing, and continues with model-specific processes. The overall analysis considers two types of modeling applications: traditional statistical models and machine learning models. The training and evaluation are conducted for both categories before performance of the models across categories is compared through statistical testing. This structure frames a logical and consistent process for evaluating model effectiveness in temperature forecasting.

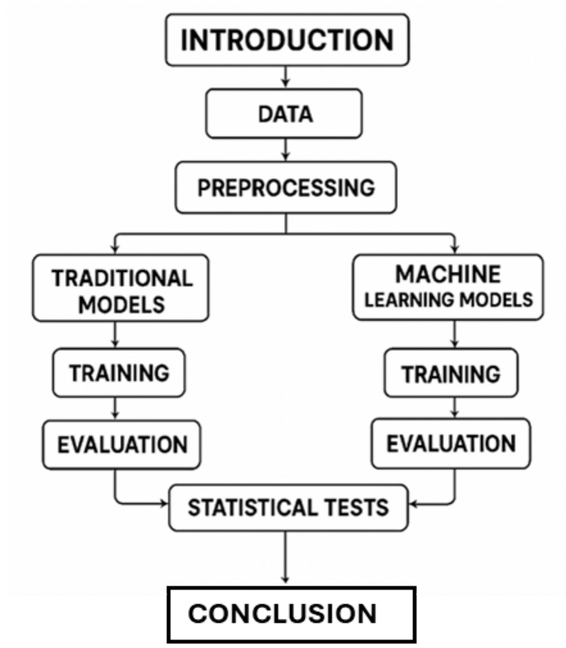

Fig 1. Structure of the Paper

3 Results and Discussion

3.1 Traditional Models

The ARIMA and SARIMA graphs from 2017 to 2024 demonstrate that SARIMA provides higher-quality forecasts with respect to actual temperatures and trends, especially when observing seasonal changes. ARIMA seems to perform relatively well in stable years, however, deviate significantly in years 2023 and 2024. In general, the visualisation of these graphs suggest that SARIMA better captures atmospheric seasonality and variability compared to ARIMA.

The graphs of actual versus predicted temperature for the years 2017 to 2024 allow for a visible comparison on the forecasting ability for two models, ARIMA and SARIMA. The graphs enable a comparison of the model predictions versus observed temperature over a temporal domain. The graphs indicate that although both models track the lower temperature trend, it is evident that SARIMA has improved depicted fit to observed temperature data, especially with seasonal variations. In the case of ARIMA the gap between the respective predicted and observed values becomes even more pronounced in visually. This is especially noticeable for 2023 and 2024, where the trend extends outward beyond the variability of the data, as in the case of both SARIMA and ARIMA ending in 2020. More specifically, ARIMA has difficulties in trying to capture variability or seasonality during the seasonal fluctuations, as seen obviously when looking at both models. The graphs visually illustrate the difference in predictive accuracy between SARIMA and ARIMA approaches.

3.2 Machine Learning Models

Machine learning models can accurately learn trends among historic sets of weather data and make educated predictions based on those trends. Unlike traditional statistical methods, machine learning models are particularly beneficial because they can learn complex, non-linear relationships in the data, and models can adjust to changing conditions if the relationships change. We

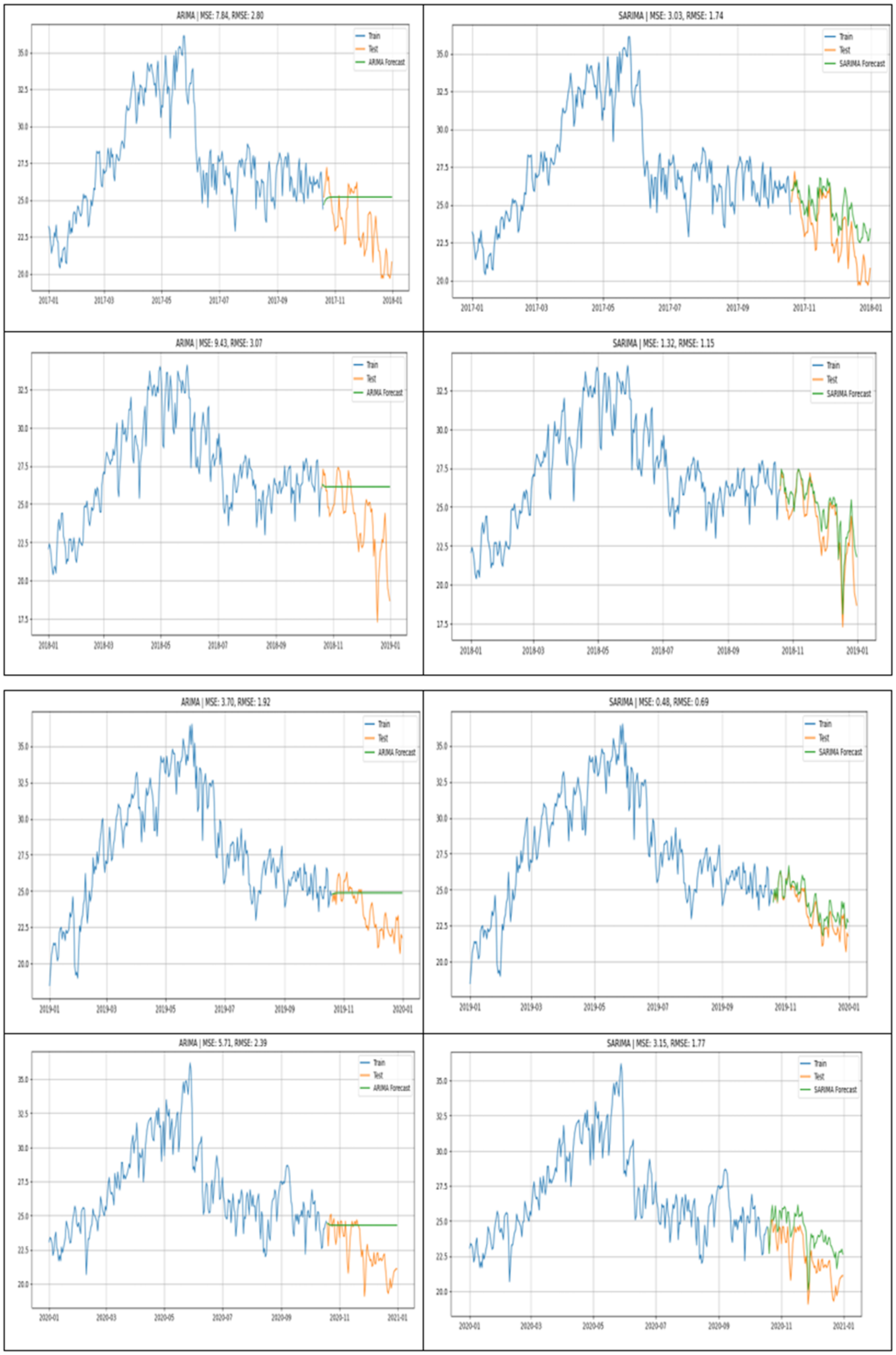

Fig 2. ARIMA and SARIMA models for the years 2017-2024 respectively

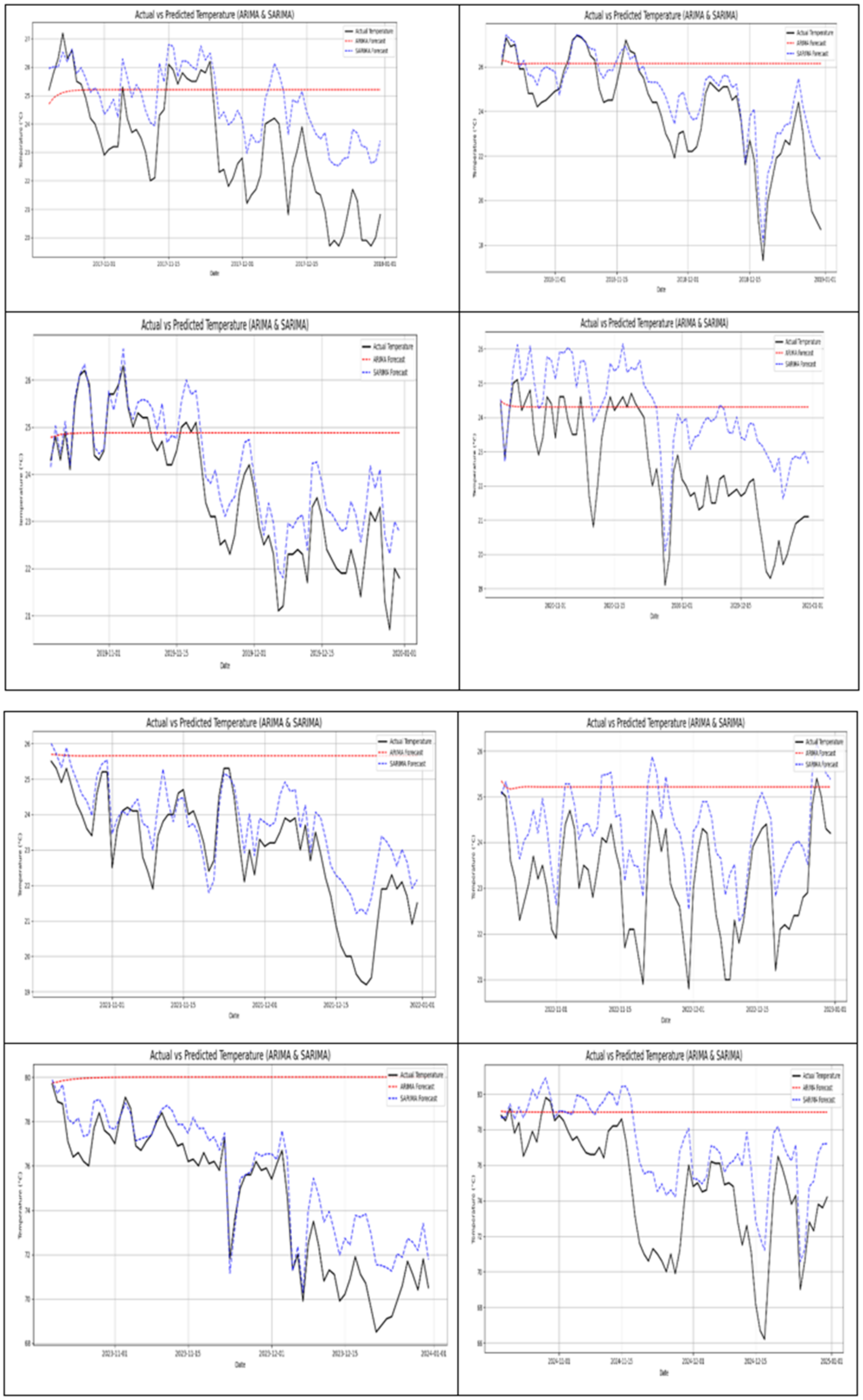

Fig 3. Actual vs predicted temperature using ARIMA and SARIMA models for the years 2017 to 2024 respectively

previously mentioned models such as Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN), and Long Short-Term Memory (LSTM)

3.2.1. Multiple Linear Regression:

The general Multiple Linear Regression equation is represented as

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_0 X_k + \varepsilon$

Where Y is the Dependent Variable

$X_1, X_2, ... X_k$ are the Independent variables

β_0 is the Y – Intercept

$\beta 1, \beta_2, ... \beta_k$ are the coefficients of each independent variable

3.2.2. Decision Tree:

A decision tree is a flowchart-like model of choices and consequences for better decision-making, and within machine learning, a supervised algorithm for classification and regression problems that has a structure that consists of nodes and branches.

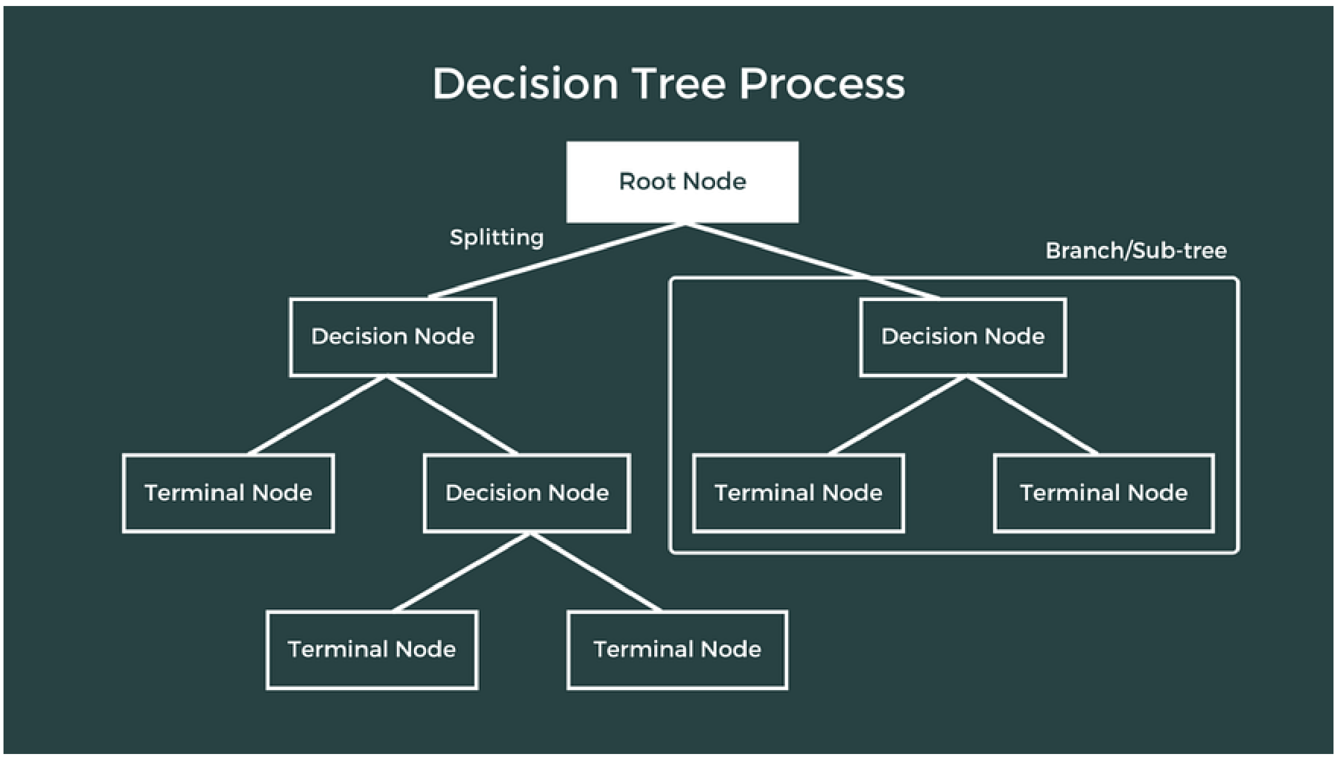

Fig 4. Decision Tree

3.2.3. Random Forest:

The Random Forest Algorithm is a strong machine learning algorithm that pulls together multiple decision trees to make high-accuracy predictions through ensemble learning.

3.2.4. Artificial Neural Networks:

Artificial Neural Networks (ANNs) are machine learning models based on neural architectures, specifically the structure and function of the human brain. ANNs are powerful computers that can be trained to perform pattern recognition tasks or make predictions, among other applications. ANNs are collections of interconnected nodes or neurons that are arranged in layers (input layer, hidden layers and output layer) and are capable of learning indirectly from data. The interconnections or weights are adjusted, and the degree of adjustment is important to behave according to some desired output.

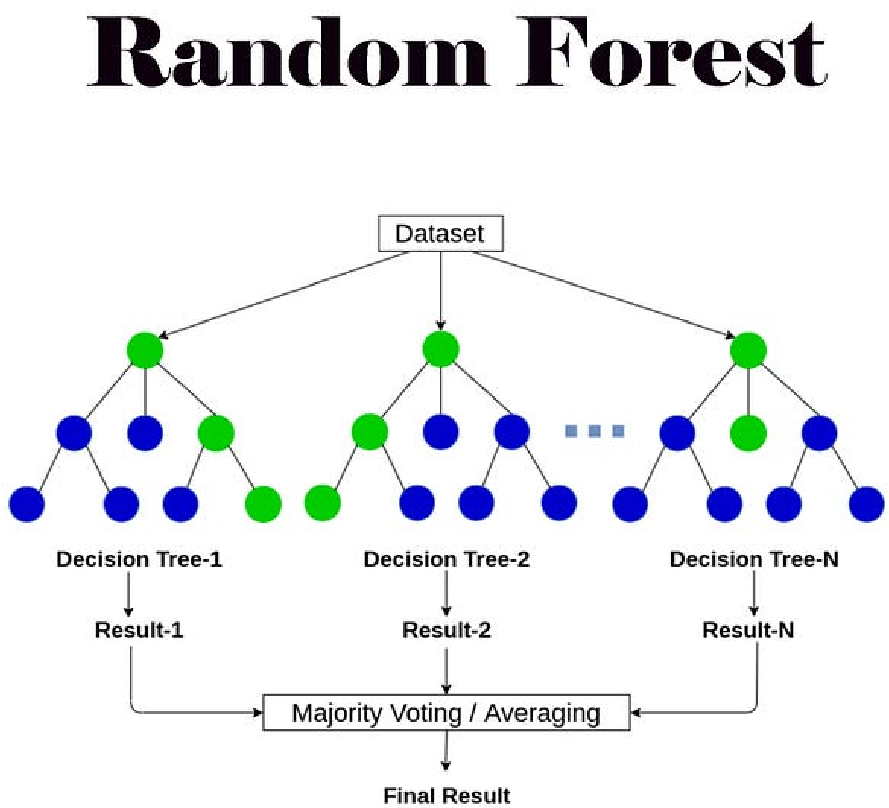

Fig 5. Random Forest

3.2.5. Long Short-Term Memory (LSTM):

LSTM is a specific type of Recurrent Neural Network (RNN) that can learn short- and long-range dependencies in sequence data. LSTMs learn both short- and long-term dependencies using memory cells or gates (input, forget, output) to decide what information to keep or forget again to easier handle sequential data, which makes LSTM well suited for developing forecasting models for time series data.

3.3 Traditional Models vs Machine Learning Models

The table below shows values of Root Mean Square Error (RMSE) between different models of temperature prediction between years 2017 and 2024. It compares two types of models in terms of RMSE, the first one being traditional, time series models like ARIMA and SARIMA, to a set of machine learning models like Multiple Linear Regression (MLR), Decision Tree (DT), Random Forest (RF), Artificial Neural Network (ANN), and Long Short Term Memory (LSTM). It is important to note that the lower the RMSE the better the model, and this is reflected in the table below. Generally, and in most years in the table, machine learning models outperformed the traditional ones.

Both ARIMA and SARIMA were tested against temperature forecasting from 2017–2024 through RMSE values. In all years, SARIMA had lower RMSE values compared to ARIMA and thus more accurately forecasted temperatures. For example, in 2018 the RMSE for ARIMA was 3.07, while the RMSE for SARIMA was only 1.15—moreover, the greatest difference was observed in 2023 (6.33 in 2023 for ARIMA compared to 1.45 in 2023 for SARIMA). SARIMA forecasted temperatures more accurately than ARIMA in 2024 as well (2.90 for SARIMA and 5.32 for ARIMA). Overall, SARIMA facilitated seasonal processes better than ARIMA did.

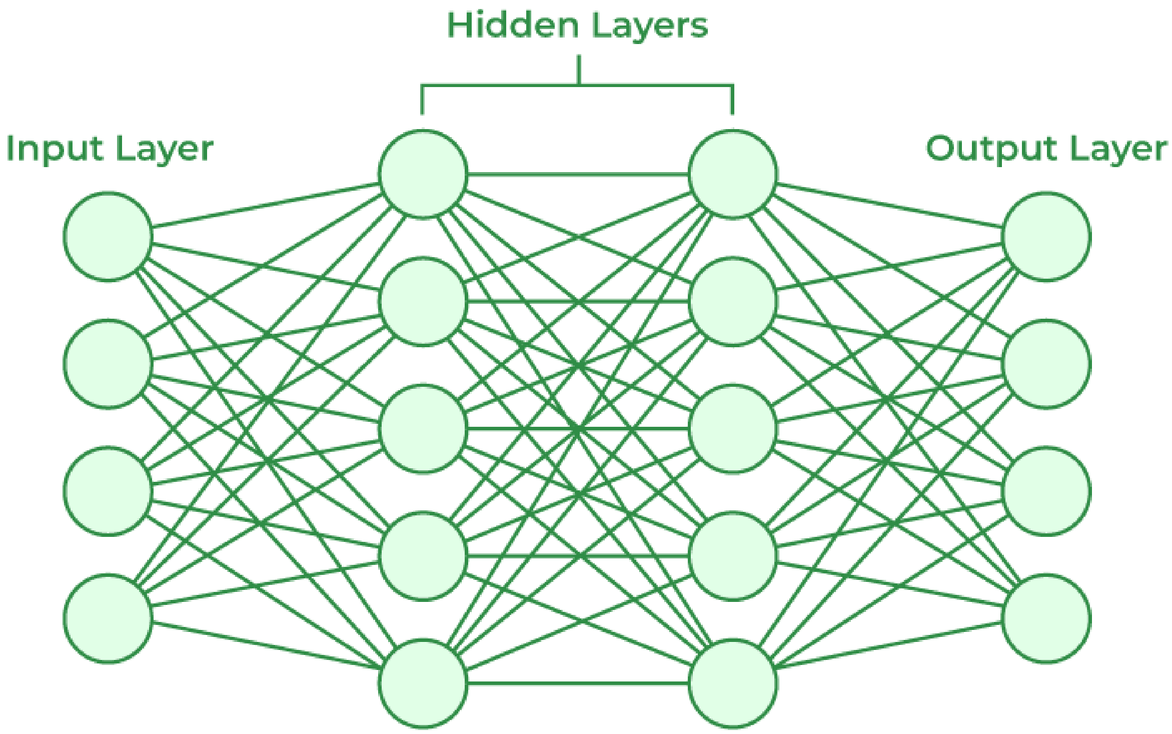

Fig 6. ANN

Table 1. Year-wise RMSE Values of Traditional and Machine Learning Models for Temperature Prediction (2017–2024)

Year	Traditional Models		ML Models									
	ARIMA	SARIMA	MLR		DT		RF		ANN		LSTM	
	RMSE	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
2017	2.8	1.74	0.9749	0.5754	0.9475	0.8322	0.9729	0.5977	0.9426	0.9523	-1.4817	1.4817
2018	3.07	1.15	0.981	0.4553	0.9347	0.8435	0.964	0.6267	0.9179	0.9463	0.2049	2.0127
2019	1.92	0.69	0.9863	0.4455	0.9568	0.7914	0.9778	0.5674	0.9633	0.7297	0.2919	1.2194
2020	2.39	1.77	0.9918	0.2921	0.9424	0.7733	0.9786	0.4710	0.9635	0.6159	0.0307	1.5961
2021	3.06	0.94	0.9822	0.3439	0.9376	0.6439	0.9616	0.505	0.8658	0.9446	-0.162	1.6687
2022	2.33	1.23	0.9819	0.4246	0.9374	0.7896	0.9609	0.6239	0.9278	0.8477	-1.3939	1.6766
2023	6.33	1.45	0.9801	0.7239	0.9537	1.1048	0.9718	0.8623	0.6239	3.1496	0.3629	2.4887
2024	5.32	2.9	0.989	0.4246	0.9374	0.7896	0.9609	0.6239	0.9329	0.8174	-1.535	1.7253

The R^2 results from the years 2017–2024 show that Multiple Linear Regression (MLR) is the best predictor, based on all years being in the 0.97 to 0.99 range, all being stable and having very high R^2 meanings, which suggests that MLR explained almost all of the variance in temperature data from year to year. Artificial Neural Networks (ANN) and Random Forests (RF) performed fairly solid as well, but the results did have fluctuations in R^2 values. Decision Trees (DT) had low results, below 0.5 in some years. The Long Short-Term Memory (LSTM) models did the worst, having mostly negative R^2 values, and therefore, may have no predictive ability at all. In conclusion, MLR model was the most consistent and most accurate option, while ANN and RF were two solid secondary options, and DT and LSTM were insufficient for the data sampled.

The machine learning models—multiple linear regression (MLR), decision tree (DT), random forest (RF), artificial neural network (ANN), and long short-term memory (LSTM) were compared year-wise from 2017 to 2024 based on RMSE values to assess their accuracy in predicting temperature. In 2017, MLR had the least RMSE at 0.5754, followed by RF (0.5977), and LSTM had the highest error at 1.4817. The same was true in 2018, where MLR had the least RMSE again at 0.4553, and LSTM again had the highest RMSE at 2.0127. MLR was again the most accurate in 2019, where it had the least RMSE at 0.4455, and LSTM again had higher error at 1.2194. In 2020, MLR was the best predictor at 0.2921 RMSE, while LSTM had 1.5961. In 2021, MLR once again had the lowest error (0.3439), while LSTM had the highest at 1.6687. For 2022, MLR had the least RMSE at 0.4246, while LSTM had 1.6766.

An exception occurred in 2023, where ANN had a severe RMSE spike to 3.1496, LSTM had a RMSE of 2.4887, and MLR was still the best, but had a higher error than previous years of 0.7239. Then in 2024, MLR improved to a better RMSE of 0.4246 and retained the top model position, while LSTM ended with 1.7253 RMSE. In summary, over the years, MLR had the lowest consistency of RMSE and predictions ability of all models, even though RF was a close second best usually, while LSTM and ANN had higher yet inconsistent levels of RMSE suggesting that simpler models like MLR and RF were more stable given the nature of predictive performance across temperature.

In 2017, SARIMA received an RMSE of 1.74 and ARIMA received 2.80 where both were substantially higher than the best performing ML model, that being MLR, with an RMSE of 0.5754 as well as validating ML was performing better. In 2018, SARIMA received 1.15 and ARIMA received 3.07 with MLR receiving an RMSE of 0.4553 validating ML was performing significantly better than SARIMA and ARIMA. In 2019, SARIMA received 0.69 which was marginally better than ARIMA at 1.92 but WAY behind MLR at 0.4455, this trend continued in 2020 where SARIMA had an RMSE of 1.77 and ARIMA an RMSE of 2.39 while MLR had an RMSE of 0.2921 - thus showing MLR performing 2 to 3 times better than SARIMA and ARIMA. In 2021, SARIMA received 0.94 and ARIMA received 3.06 or MLR received a much less RMSE of 0.3439. In 2022, SARIMA and ARIMA had 1.23 and 2.33 respectively again, behind MLR at 0.4246. In the following year, 2023, ARIMA spiked to 6.33 so SARIMA was at 1.45 with MLR significantly better than either at an RMSE of 0.7239 but with an overall increase. In the most recent data for 2024, ARIMA had an RMSE of 5.32 and SARIMA had an RMSE of 2.90 while MLR had an RMSE of 0.4246 and still performed the best overall and although less RMSE than the previous seasons, this trend has continued.

3.4 Independent Two-Sample t-Test

The two-sample t-test for equal variances was conducted to compare ARIMA and SARIMA. The two groups were tested on the means and variances to determine whether the observed difference was statistically significant.

Null Hypothesis H_0: The means of RMSE Values of ARIMA and SARIMA are equal.

Alternative Hypothesis H1: The means of RMSE values of ARIMA and SARIMA are not equal.

Level of Significance: α = 5% = 0.05.

Table 2. t-Test: Two-Sample Assuming Equal Variances

	ARIMA	SARIMA
Mean	3.4025	1.48375
Variance	2.45874	0.46514
Observations	8	8
Pooled Variance	1.46194	
Hypothesized Mean Difference	0	
df	14	
t Stat	3.17383	
P(T<=t) one-tail	0.00338	
t Critical one-tail	1.76131	
P(T<=t) two-tail	0.00676	
t Critical two-tail	2.14479	

The two-sample t-test with equal variance compares ARIMA (mean = 3.4025) to SARIMA (mean = 1.4838). The calculated t Stat = 3.1738 exceeds the t critical two-tail = 2.1448, and the p-value (two-tail) = .0068, which is less than the typical significance level of 0.05. There is a statistically significant difference in means between the two variables. For this reason, we reject the null hypothesis of equal means and conclude that ARIMA has a higher mean than SARIMA.

3.5 ANOVA One – Way Classification

To assess the performance of different models, a one-way ANOVA test was utilized. This calculation determines the significance of the difference between mean values from multiple groups. This test helps to assess if at least one of the models performs in a different manner than the others.

Null Hypothesis H_0: All group means are equal.

i.e., $\mu_{MLR} = \mu_{DT} = \mu_{RF} = \mu_{ANN} = \mu_{LSTM}$

Alternative Hypothesis H_1: At least one of the group mean is significantly different from others.

Level of Significance: α = 5% = 0.05.

Table 3. ANOVA: Single Factor

SUMMARY						
Groups	**Count**	**Sum**	**Average**	**Variance**		
MLR	8	3.6853	0.46066	0.0182		
DT	8	6.5683	0.82104	0.01681		
RF	8	4.8779	0.60974	0.01381		
ANN	8	9.0035	1.12544	0.6829		
LSTM	8	13.8692	1.73365	0.1432		
ANOVA						
Source of Variation	**SS**	**df**	**MS**	**F**	**P-value**	**F crit**
Between Groups	8.13397	4	2.03349	11.6213	4.15E-06	2.64147
Within Groups	6.12429	35	0.17498			
Total	14.2583	39				

The one-way ANOVA analyzed the means of five groups (MLR, DT, RF, ANN, and LSTM). The F value is 11.62, which exceeds the F critical value of 2.64. The p-value (4.15E-06) is also significantly smaller than 0.05. This means that there are meaningful differences between the means of the groups, thus at least one model's performance (mean value) is statistically different than the others.

Similarly, the relative performance of traditional statistical models in comparison to modern machine learning techniques in our study is similar to the burgeoning set of literature across various fields. Our analysis indicates that machine learning models, particularly Multiple Linear Regression (MLR) and Random Forest (RF), performed better in our context than traditional models like ARIMA and SARIMA, respectively, which was indicated by lower RMSE values across all observed years of temperature data for Hyderabad. Our results find parallel evidence in [(12)] who indicated RF had a lower RMSE (24.88) than ARIMA (31.71) and exponential smoothing (34.95) for industrial demand forecasting, significantly better results than traditional statistical model performance. [(5)] Papers evaluate traditional models which include ARIMA together with SARIMA and compare these with advanced machine learning techniques for time series prediction. The studies find that machine learning models particularly ensemble models such as Random Forest and REACH perform better than ARIMA in terms of predictive accuracy. In healthcare applications, [(6)], [(8)] showed very modest increases in predictive accuracy by machine learning models as compared to logistic regression; however, in the context of weather forecasting, we observed more of a separation in performance. In fact, our observations are corroborated by [(14)], who reported success of ML and deep learning models in non-linear multivariate time series contexts, which are less effective for traditional models. In addition, [(2)] showed that a hybrid CNN–LSTM model outperformed ARIMA for hourly temperature forecasting (with RMSE values of 1.38°C versus 1.56°C). In our analysis, LSTM did not perform quite well as we expected, likely due to the nature of daily temperature data in our analysis and the common lack of deep temporal dependencies, which recurrent neural networks would use to their advantage. Several recent papers provide evidence that there is increased acceptance of ML models across applications. For example, [(10)] show that machine learning models had AUCs of 0.819 for diabetes prediction, which is higher than the traditional model used. Likewise, [(13)] showed that ML models were superior on c-index (0.913 vs 0.835) outcomes for predicting heart failure. Outcomes such as these demonstrate that ML is becoming more commonly applicable even in critical domains. Both [(1)], [(9)] emphasize that although ML models do typically produce higher predictive performance (up to 25% higher accuracy in [(1)]), traditional methods have the benefit of explicability and ease of deployment, something that remains a consideration in public-facing forecasts. Our results also match [(7)], [(11)] who found that CNNs and other deep learning models outperformed traditional machine learning in facial recognition and medical imaging respectively. However as in our study model complexity doesn't always guarantee better performance. For example, MLR a simple model always gave the lowest RMSE in our dataset, so simplicity can still give high accuracy when data is favorable. Finally [8], [13] advise not to use ML without considering context. Our results support this view—while ML outperformed ARIMA and SARIMA in our case, interpretability, resource requirements and overfitting risk are still important to consider especially with models like ANN and LSTM.

This study, in a way, launches directly from our earlier work published in the Indian Journal of Science and Technology [15], which dealt with the performance of machine learning algorithms, such as MLR, Decision Tree, Random Forest, ANN, and LSTM, for temperature forecasting under multivariate weather data with inputs like humidity, solar radiation, wind speed, and precipitation. That study found that MLR outclassed virtually all the others throughout, reaching a very high R^2 of 0.989 with an RMSE as low as 0.4246 in 2024, whereas LSTM showed erratic performances. By contrast, what we presently do is univariate modeling using the temperature as the single input variable and extending the comparison of models to include standard time series models, namely ARIMA and SARIMA, along with the same machine learning techniques identified above. The results from the present study evince that SARIMA, owing to its capacity to capture seasonality, is generally better than ARIMA but is still outperformed by MLR and Random Forest over all the years. This disparity between the two studies showcases the strength of machine learning methods, whether in present in multivariate or univariate scenarios and point to the unfinished-theoretical challenges classical time series methods undertake to model temperature data with any efficient degree of complexity, even when the d-energy of time dependencies is considered. Consequently, while the [15] emphasized the importance of enhanced climate modeling with richer features, this paper shows that the machine learning models can maintain their dominance in predictability and consistency even in much simpler data arrangements.

Random Forest apparently does the trick for temperature prediction across Indian cities as Alone et al. (2025) report. Contrarily, in our study, RF may be good, but MLR has bested all the models, including RF, in univariate forecasting of temperature for Hyderabad. Our work adds value by extending the comparison of RF with traditional and other ML models year-wise.

4 Conclusion

The t-test outcomes demonstrate that the two variables (ARIMA & SARIMA) involved do differ in mean. In other words, the variables do not operate at the same level or field, with one variable consistently performing better than the other. Conversely, the ANOVA test indicates that the multiple groups do differ and therefore not all models operate at the same level. Each performs differently, with at least one completing the task differently from the others. Overall, both tests confirm that both pairwise and group-wise comparisons are meaningful.

The review of RMSE values from 2017 to 2024 reveals that machine learning models showed higher predictive power than traditional models for temperature prediction. The MLR model had the lowest RMSE value of all models in all years, with values ranging from 0.2921 (2020) to 0.7239 (2023). The best traditional model, SARIMA, has higher RMSE values than MLR, ranging between 0.69 (2019) to 2.90 (2024); ARIMA had the highest RMSE errors overall, ranging from 2.81 Peaking at 6.33 in 2023.

Random Forest (RF) performed well of the machine-learning models, scoring an RMSE of between 0.4710 and 0.8623, and usually scored second to MLR. Decision Tree (DT) was moderately successful, scoring an RMSE of between 0.6439 and 1.1048. Artificial Neural Network (ANN) and Long Short-Term Memory (LSTM) were less consistent with their accuracy. ANN achieved relatively low RMSE in some years (i.e., 0.6159 in 2020) but then scored very high RMSE in others (i.e., 3.1496 in 2023). LSTM, despite being the only model that accounts for sequential datasets, consistently yielded higher RMSE scores ranging from 1.2194 to 2.4887, suggesting a less adequate method for use with this type of dataset.

In terms of R^2 values, the Multiple Linear Regression (MLR) model consistently performed best, explaining roughly 97%–99% of the variance. Both ANN and RF performed moderately well, demonstrating some fluctuations. The DT model performed more poorly than ANN and RF, but better than LSTM. The LSTM model performed the worst, demonstrating mostly negative R^2 values, indicating poor predictive ability.

To sum it up, MLR was the best model and the one that stood out to be very stable, with RF being the second best. This indicates that simpler machine learning methods work significantly better than time series models such as ARIMA and SARIMA when estimating yearly variation in temperature data, with year-to-year variation being somewhat consistently extrapolated.

This performance can be attributed to the fact that the temperature data showed relatively linear relationships over time and not very complicated relationships between features. MLR was able to replicate the underlying trends and seasonal changes without overfitting, which tends to occur in more complicated models especially when the dataset isn't large or particularly varied. In addition, MLR, like other simple models, had a lower number of parameters that require far less computational power, resulting in a level of stability and interpretability that will positively impact decision making. MLR also had a reasonable RMSE in the most complicated and noisy years, whereas LSTM and ANN (while having a representation of the temperature data in potential algorithms) needed a proportionately large dataset in order to provide a sensitized generalizability in forecasting that allowed the model to benefit compared to MLR. MLR proved to be a simple, representative, and realistic model for forecasting real-world, time series urban temperature data from Hyderabad.

4.1 Limitations:

- Because the study only uses data from IMD Hyderabad, it might not apply to other areas.
- Other weather factors were excluded; only temperature was used.
- Long-term performance was not assessed; short-term forecasting was the main focus.

4.2 Future Scope:

Model parameter optimization together with the inclusion of extra environmental variables will help improve the prediction accuracy of future studies. Researchers should investigate the performance of CNN-LSTM models alongside Transformer models to detect better temporal patterns. Adding multiple regional data points to the analysis enables spatial forecasting capabilities while real-time integration of models improves climate monitoring and early warning capabilities. The models should receive modifications to produce extended forecast periods which will enable better examination of climate change effects.

5 Acknowledgments

The author would like to thank the India Meteorological Department, Begumpet, Hyderabad, Telangana.

- Author Contributions Statement:
  - Author 1,2: Basic statistical analysis, writing, Analysis and editing the documentation
  - Author 3: Guidance
  - Author 4: Conceptualization & Methodology
- Conflict of statement: The authors declare no conflicts of interest.
- Ethical Compliance statement: This study does not involve human participants.
- Funding Statement: This research does not support any funding agency.
- Data Availability Statement: The dataset was taken from the India Meteorological Department (IMD), Hyderabad.

References

1) Alanazi BS. A Comparative Study of Traditional Statistical Methods and Machine Learning Techniques for Improved Predictive Models. *International Journal of Analysis and Applications*. 2025;23:18–18. https://doi.org/10.28924/2291-8639-23-2025-18.
2) Saa ED, Ranathunga L. Comparison between arima and deep learning models for temperature forecasting. *arXiv preprint arXiv:201104452*. 2020. https://doi.org/10.48550/arXiv.2011.04452.
3) Kumar R, Gupta M, Agarwal A, Mukherjee A, Islam SM. Epidemic efficacy of Covid-19 vaccination against Omicron: An innovative approach using enhanced residual recurrent neural network. *Plos one*. 2023;18(3):e0280026. https://doi.org/10.1371/journal.pone.0280026.
4) Wu CW, Shen HL, Lu CJ, Chen SH, Chen HY. Comparison of different machine learning classifiers for glaucoma diagnosis based on spectralis OCT. *Diagnostics*. 2021;11(9):1718. https://doi.org/10.3390/diagnostics11091718.
5) Poonawala-Lohani N, Riddle P, Adnan M, Wicker J. A novel approach for time series forecasting of influenza-like illness using a regression chain method. *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2022*. 2021;p. 301–312.
6) Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of machine learning methods with traditional models for use of administrative claims with electronic medical records to predict heart failure outcomes. *JAMA network open*. 2020;3(1):e1918962–e1918962. https://doi.org/10.1001/jamanetworkopen.2019.18962.
7) Caroppo A, Leone A, Siciliano P. Comparison between deep learning models and traditional machine learning approaches for facial expression recognition in ageing adults. *Journal of Computer Science and Technology*. 2020;35:1127–1146. https://doi.org/10.1007/s11390-020-9665-4.
8) Zeng J, Zhang D, Lin S, Su X, Wang P, Zhao Y, et al. Comparative analysis of machine learning vs. traditional modeling approaches for predicting in-hospital mortality after cardiac surgery: temporal and spatial external validation based on a nationwide cardiac surgery registry. *European Heart Journal-Quality of Care and Clinical Outcomes*. 2024;10(2):121–131. https://doi.org/10.1093/ehjqcco/qcad028.
9) Bennett M, Kleczyk EJ, Hayes K, Mehta R. Evaluating similarities and differences between machine learning and traditional statistical modeling in healthcare analytics. IntechOpen. 2022. https://doi.org/10.5772/intechopen.105116.
10) Choi SG, Oh M, Park DH, Lee B, Lee YH, Jee SH, et al. Comparisons of the prediction models for undiagnosed diabetes between machine learning versus traditional statistical methods. *Scientific reports*. 2023;13(1):13101.
11) Wan KW, Wong CH, Ip HF, Fan D, Yuen PL, Fong HY, et al. Evaluation of the performance of traditional machine learning algorithms, convolutional neural network and AutoML Vision in ultrasound breast lesions classification: a comparative study. *Quantitative imaging in medicine and surgery*. 2021;11(4):1381. https://doi.org/10.21037/qims-20-922.

12) Quevedo FCS. A Comparison of Machine Learning and Traditional Demand Forecasting Methods. *Master's thesis, Clemson University*. 2020. Available from: https://tigerprints.clemson.edu/all_theses/3367.
13) Shin S, Austin PC, Ross HJ, Abdel-Qadir H, Freitas C, Tomlinson G, et al. Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality. *ESC heart failure*. 2021;8(1):106–115. https://doi.org/10.1002/ehf2.13073.
14) Jr MEB, McDonald J, Toth T, Sadler B. Traditional vs Machine Learning Approaches: A Comparison of Time Series Modeling Methods. *SMU Data Science Review*. 2023;7(2):2.
15) Dhwani N, Pranay P, Haragopal VV, Manjusha T. Year-wise Longitudinal Comparative Analysis of Temperature Prediction Models for Hyderabad (2017–2024). *Indian Journal of Science and Technology*. 2025;18(27):2179–2184. https://doi.org/10.17485/IJST/v18i27.559.