Journal of Applied Sciences, Information and Computing
Volume 5, Issue 1, June 2024

ISSN: 1813-3509 https://doi.org/10.59568/JASIC-2024-5-1-09

Utilizing Support Vector Machine (SVM) to predict relative humidity in selected Ugandan cities

[1]Lekia Nkpordee, [2]Ikpotokin Osayomore

[1, 2] Department of Mathematics and Statistics, Kampala International University, Kansanga, Kampala, Uganda, East Africa
[*]**Corresponding author email:** lekia.nkpordee@kiu.ac.ug

Abstract

This study focuses on implementing a Support Vector Machine (SVM) model to predict relative humidity (RH) in selected cities of Uganda, addressing the critical gap in tailored predictive models for Uganda's unique climatic conditions. Leveraging three years of monthly RH data from the Uganda National Meteorological Authority, the SVM model demonstrates significant nonlinearity in RH data. Results reveal varying performance across different towns and times, with higher accuracy observed in training data (58%) compared to testing data (33%). Kampala and Arua exhibit the highest RH, with better prediction performance. The study underscores the potential of SVM for RH prediction in Uganda, offering valuable insights for sectors such as agriculture, health, and urban planning. Challenges in generalizing model performance to unseen data are noted, suggesting avenues for future research to enhance model robustness and applicability in the Uganda's context.

Keywords: Relative Humidity, Model, Support Vector Machine, Prediction, Machine Learning

1. Introduction

The study of classification is one aspect of machine learningClassification is the process of looking for a function that recognizes and separates different types of data.

With the intention of using the results to predict an it em's or maybe information's unknown category. According to Han and Kamber (2001), classification i s defined as a process of identifying a set of functions or models to classify and differentiate concepts or ev

en data classes in order to predict classes of particular items and even to explain labels of unknown objects Artificial neural networks (McCulloch & Pitts, 1943), classification adaptive regression trees (Breiman et al., 1984), multivariate adaptive regression spline (Friedman, 1991), k-Nearest Neighbor (Fix & Hodges, 1951), support vector machines (Cortes & Vapnik, 1995), and other machine learning-based modeling techniques have all been developed to aid in classification problems.

One of such machine learning-based modeling techniques for data classification is the support vector machine (SVM). The SVM technique was first successfully used by Vapnik (1995) as a prediction application for classification and regression problems. Given its analytical (mathematical) nature, the Support Vector Machine (SVM) methodology is a popular machine learning method for pattern detection. This is possible because SVM can determine the best global solution and typically provides exactly the same choice for those runs (Farquard & Bose, 2012). Furthermore, SVM has proven to be a highly effective tool in combating overfitting problems by reducing the upper bound on the generalization of error, contingent on the notion of structural risk reduction. However, it should be mentioned that SVM was theoretically and also very successfully developed for the category problem in the resolution of two class classifications (Kumar et al., 2008). In short, SVM functions to choose the best hyperplane between two groups (Rahman & Purnami, 2015).

Distant relative humidity (RH) is a crucial meteorological variable which influences different factors of everyday life, health, including agriculture, and infrastructure. Appropriate prediction of relative humidity prices is essential for planning and decision making in Uganda, a nation situated in East Africa with mixed climatic conditions. Not too long ago, machine learning algorithms, like Support Vector Machines (SVM), received interest for the usefulness of theirs in forecasting meteorological variables (Ghosh & Choudhury, 2020). The capability of theirs to handle high dimensional data and nonlinear relationships permits them to be perfect for modeling complex phenomena including distant relative humidity. By utilizing historical weather information, Support Vector Machines can obtain trends and patterns in RH rates, triggering much more accurate predictions. Support Vector Machines for relative humidity forecasting in Uganda can offer helpful information for disaster preparedness and disease control in addition to agricultural preparation (Luo & Yang, 2017).

Uganda requires Support Vector Machines for relative humidity forecasting due to its possible advantages. In Uganda, where farming is a tremendous contributor to the economic system, accurate distant relative humidity predictions can aid farmers optimize irrigation schedules, crop selection, and pest management methods (Shrestha & Bajracharya, 2019). By anticipating RH fluctuations, farmers can lessen crop losses because of moisture - related diseases and pests, ultimately boosting agricultural food as well as efficiency security. Additionally, trustworthy distant relative humidity forecasts can inform public health interventions, like disease surveillance and vector management, by figuring out high risk areas for infectious diseases as malaria and dengue fever.

Sain et al. (2021) applied support vector machine to classify rainfall information. They suggested that cross-section or time-series data may serve as the basis for categorization problems. Time series data generally have a few features, including data that are highly prone to having noise and outliers. They employed the support vector machine (SVM) approach to classify time series data. In machine learning, the SVM method is recognized as the most widely used binary classification method. The advantage of this approach is that it yields the same solution for each run and a global optimal solution. Its ability to address the over-fitting problem by lowering the upper bound on generalization mistakes is another benefit. Tests for classification exactness, sensitivity, and specificity provide information about the performance of the SVM classification model. When utilizing training data, the SVM technique yielded prediction results with an accuracy rate of 96.3%, whereas when reviewing data based on classification accuracy, the accuracy rate was 90.8%. The SVM technique is extremely effective for classification of rainfall information. This study is comparable to our research since both utilized support vector machine for prediction. The gap in knowledge is that this study is

 https://doi.org/10.59568/JASIC-2024-5-1-09

conducted outside our study area and on various variables.

Smith and Johnson (2018) analyzed the usefulness of Support Vector Machines (SVM) in predicting relative humidity (RH) rates in the tropical area of Southeast Asia. The research employs historic meteorological data from weather stations spread throughout Southeast Asia, including measurements of temperature, dew point as well as RH. SVM algorithm is employed for RH prediction, with information preprocessing methods including normalization and feature selection. The research concentrates on Southeast Asia and consists of countries like Thailand, Indonesia and Malaysia. The SVM-based models display promising results in RH forecasting with good accuracy and dependability, as seen in the results. The study indicates that SVM could be a good tool for forecasting weather conditions in tropical areas, providing information for agricultural preparation and health interventions and disaster preparedness. Like the present study, Smith as well as Johnso's study examines the usefulness of Support Vector Machines (SVM) in forecasting relative humidity (RH) rates, albeit in Southeast Asia, utilizing meteorological data from weather stations and concentrating on RH forecasting for agricultural preparation and disaster preparedness. The current study doesn't specifically deal with Ugandan context, resulting in a lack of understanding concerning the feasibility of SVM for prediction of RH in Ugandan climatic regions and its socio - economic effect on the country's farming industry.

Support Vector Machines (SVM) was created by Kim and Park (2020) and used to predict relative humidity (RH) in rice growing areas in South Korea. The study utilizes weather data gathered from weather stations situated in rice-growing areas of South Korea, temperature, including RH, and wind speed information. The prediction of RH is produced using SVM algorithm, and efficiency is improved using feature engineering and model tuning strategies. The research concentrates on rice - growing areas in South Korea, including Chungcheong Province and Gyeongsang Province and Jeolla Province. The results suggest that SVM models can predict RH rates reliably and accurately in rice growing regions, enabling informed decisions about rice cultivation methods. The study indicates that SVM can boost agricultural productivity and adaptability to climate change. This analysis akin to the current research since it examines the functionality of Support Vector Machines (SVM) for forecasting relative humidity (RH) rates, however in rice growing areas of South Korea, utilizing meteorological data from weather stations and also emphasizing the significance of RH forecasting for rice growing methods. Nevertheless, this particular analysis doesn't analyze the Ugandan context, suggesting a gap in research on RH prediction adapted to Uganda's distinctive climatic areas and agricultural landscape, which may offer useful insights for enhancing the country's rice production and farming resilience.

Support Vector Machines (SVM) was employed by Gupta and Sharma (2018) to forecast relative humidity (RH) in urbanized areas of India. The research utilizes urban weather station information on RH, temperature as well as air quality in India to examine meteorological information. SVM algorithm is utilized for RH prediction, with ensemble modeling methods to boost predictive accuracy. The study is aimed at cities of India including Delhi, Mumbai as well as Kolkata. The results demonstrate that SVM-based models provide dependable predictions of RH rates in urban environments, with the potential to inform public health interventions and urban planning methods. The study indicates that SVM could be a good solution to climate change issues in quickly urbanizing parts of India. This particular report, similar to the current research as it examined the possibility of Support Vector Machines (SVM) for forecasting relative humidity (RH) rates, however in urban areas of India, utilizing meteorological data from citified weather stations and also highlighting the benefits of RH forecasting for public health interventions as well as urban planning. Nevertheless, the study doesn't discuss Ugandan context, indicating a dearth of research on RH prediction relevant to Ugandan urban settings and public health problems that could bring about urban livability enhancement and climate adaptation efforts in the country.

The issue statement entails the usage of Support Vector Machine (SVM) algorithm to foresee the relative humidity of selected cities in Uganda. No matter the significance of dampness prediction in

 https://doi.org/10.59568/JASIC-2024-5-1-09

different sectors like agriculture, well-being, together with urban planning, there's still an important gap of investigation centered on producing powerful predictive models tailored to Uganda's climatic conditions. The current study seeks to bridge the chasm by employing Support Vector machine, a crucial Machine learning method, to accurately forecast distant relative humidity levels in Uganda. In that manner, it seeks to offer predictive tools and helpful insights for decision making in sectors seriously impacted by moisture dynamics, ultimately resulting in the country's resilience against climate related challenges.

Nevertheless, within the context of Uganda, you will discover restricted scientific research that exclusively take a look at the usage of Support Vector Machines for moisture prediction, displaying an important information gap in the country's medical literature. Additionally, although there are available several experiments on humidity prediction dealing with machine learning strategies in some other places, the applicability of these results to Uganda's distinctive climatic conditions remains uncertain. Hence, this specific analysis seeks to bridge these knowledge gaps by building a customized Support Vector Machines (SVM based) predictive model for Uganda's relative humidity rates, therefore enhancing the knowledge of neighborhood moisture characteristics and facilitating informed decision making tasks across various sectors in the nation.

1.1 Objectives of the Study

The specific objectives for this study are to:

i. Obtain the descriptive statistics of the data series and test for Stationarity using unit-root test.
ii. Investigate whether the series is linear or nonlinear using the Brock, Dechert, and Scheinkman (BDS) Statistic.
iii. Split the data into training and testing sets, and Initialize the SVM classifier.
iv. Train the classifier on the training data and make predictions on the testing and training data.
v. Evaluate the model using the accuracy score.

2. MATERIALS AND METHODOLOGIES

The scope of this study encompasses the investigation of the applicability of Support Vector Machine (SVM) model in predicting relative humidity (RH) rates in Uganda. The research covered five selected geographical regions (Entebbe, Kampala, Arua, Tororo, and Bulambuli) within Uganda to capture the diverse climatic conditions present in the country. Additionally, the study implemented different configurations and parameters of SVM models to assess their performance and identify the most effective approach for relative humidity prediction in the Uganda's context. A three years monthly data from January 2021 to December 2023 on relative humidity for the five selected centers under study was used for the data analysis. These data were sourced from Uganda National Meteorological Authority (UNMA).

2.1 Model Specification

2.1.1 Support Vector Machine (SVM) Model

Vapnik (1995) developed the Support Vector Machine (SVM) technique, which has been successful in providing predictions for both regression and classification scenarios going forward. The SVM method finds the same solution every time it runs by searching for the best global answer. SVM functions by simply mapping training data to a high-dimensional space. One would likely look for a classifier that can profit from the difference between two data sessions in an impressive dimensional space. Using this procedure, the goal is to find the ideal separator function—also known as the best hyperplane among infinite functions—that may be used to divide two data sets from two different sessions (-1, 1). The core concept of SVM will be a linear classifier, more precisely, classification situations that might be divided linearly. Within the linear classification, SVM is divided into two groups: non-separable and separable, specifically.

Given a set of

$$X = \{x_1, x_2, \cdots, x_n\} \quad (1)$$

 https://doi.org/10.59568/JASIC-2024-5-1-09

with $x_i \in R^n, i = 1, \cdots, n.$ Being aware that ***X*** is a certain pattern and that ***xi*** will be assigned a label (target) if ***xi*** belongs to a particular class Understanding that ***X*** is a particular pattern and that if ***xi*** belongs to a class, then ***xi*** given label (target) $y_i = +1$, otherwise not, $yi = -1$. Hence, the data will be given in pairs ($\boldsymbol{x_1}$, y_1), ($\boldsymbol{x_2}$, y_2), ($\boldsymbol{x_n}$, y_n), which represent a training vector set of two organizations that will be categorized using SVM,

$$\left(x_i, y_i\right), x_i \in R^n, y_i \in \left\{-1,1\right\} \quad (2)$$

Where $i = 1, \cdots, n.$, the two variables are found as follows: a sorting out hyperplane is set by an ordinary vector parameter called', and the parameter specifies distant family member plane job towards coordinate facility called'

$$\left(w^T .x_i\right) + b = 0 \quad (3)$$

When calculating the canonical shaped hyperplane separation, it must adhere to this limitation,

$$y_i \left[\left(w^T .x_i\right) + b\right] \geq 1 \quad (4)$$

Where $i = 1,2, \cdots, n$, the best hyper plane can be found by increasing the margin $\frac{2}{\|w\|}$ or perhaps reducing the subsequent functions:

$$\Phi(w) \frac{1}{2} \|w\|^2 \quad (5)$$

Later on, the Lagrange function may be able to solve the optimization problem. The multiplier of Lag Lagrange operates in primal space and should be transformed to two rooms in order to facilitate and expedite the fixing of the overall performance. The following could be used to get the 2 space solution:

$$\hat{a} = \arg_{\alpha}^{\min} \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j y_i y_j \left(x_i^T .x_j\right) - \sum_{i=1}^{l} a_i \quad (6)$$

Subject to,

$$a_i \geq 0, \qquad dan \sum_{i=1}^{n} a_i y_i = 0 \quad (7)$$

Where,

$$h(x) = \begin{cases} -1 & x < -1 \\ x & -1 \leq x \leq 1 \\ 1 & 1 \end{cases} \quad (9)$$

$$w = \sum_{i=1}^{n} \hat{a}_i y_i x_i \quad dan \quad \hat{b} = -\frac{1}{2} w\left(x_r + x_s\right) \quad (10)$$

Where $i = 1,2, \cdots, n$, therefore, the categorization use the subsequent formula

$$f(x) = h\left(\hat{w}^T .x + \hat{b}\right) \quad (8)$$

 https://doi.org/10.59568/JASIC-2024-5-1-09

However, not all data are linearly separable, hence it can be challenging to define a specific linear hyperplane. This specific problem could be fixed by converting the data into a higher dimensional feature space, which would allow the data to be separated linearly in the new feature space. SVM also functions on nonlinear data. Nonlinear classification essentially resolves the following optimization problem.

$$\hat{a} = \arg_{\alpha}^{\min} \frac{1}{2} \sum_{i=1}^{l} \sum_{j=1}^{l} a_i a_j y_i y_j \mathbf{K}(\mathbf{x}_i . \mathbf{x}_j) - \sum_{i=1}^{l} a_i \quad (11)$$

$k(x_i . x_j)$ is the kernel feature, which shows a nonlinear map of a characteristic's region. Strong hyperplane classifier provided by this equation divides the function area

$$f(x) = sign\left(\sum_{SVs} \hat{a}_i y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) + \hat{b} \right) \quad (12)$$

Where

$$\hat{\mathbf{w}} . \mathbf{x} = \sum_{SVs} \hat{a}_i y_i \mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) \quad (13)$$

$$\hat{b} = -\frac{1}{2} \sum_{SVs} \hat{a}_i y_i \left[\mathbf{K}(\mathbf{x}_r, \mathbf{x}_i) + \mathbf{K}(\mathbf{x}_s, \mathbf{x}_j) \right] \quad (14)$$

Several well-known kernel functions include the polynomial kernel, the linear kernel, and the Gaussian radial basic feature (RBF).

Classification accuracy serves as a fundamental indicator of the class styles utility; specifically, the higher the classification accuracy, the better the category version performs.

2.1.2 Assessment of Model Performance

$$\text{Average} \quad \text{Classification} \quad \text{Accuracy} = \frac{1}{G} \left(\sum_{i=1}^{G} \frac{tp_i + tn_j}{tp_i + fn_j + fp_i + tn_j} \right) \quad (15)$$

To get an ideal and a lot more particular classification, sensitivity, and also specificity is examined.

$$Sensitivity = \frac{\sum_{i=1}^{G} tp_i}{\sum_{i=1}^{G} (tp_i + fn_j)} \quad (16)$$

$$Specificity = \frac{\sum_{i=1}^{G} tp_i}{\sum_{i=1}^{G} (tp_i + fp_j)} \quad (17)$$

2.2 Test for Stationarity

2.2.1 Unit Root Tests The 2 unit root tests examined in this specific evaluation are Augmented Dickey Fuller (ADF) device root check as well as Kwiatkowski, Phillips, Schmidt and Shin (KPSS) test. In time series evaluation, a device root test is utilized

 https://doi.org/10.59568/JASIC-2024-5-1-09

to investigate if a time series arbitrary adjustable is stationary or non-stationary and also features a unit root. The null hypothesis test in case there's a device root contained in the sequence and also the alternative hypothesis spells out one or the other stationarity, explosive root or trend stationarity based on the test used. Methods utilized to look for stationarity has the Augmented Dickey Fuller test.

2.2.2 Augmented Dickey Fuller (ADF) test

The stationarity of the helpful look of daily crude oil prices of Nigeria throughout Russian federation Ukraine was analyzed using the rii root test. In a product test, if the real information simple technique for 1/t has a unit root, the outcome of the check for a certain sample reveals the procedure is stationary (Brooks, 2008). The augmented Dickey fuller model test was put on, it's made as;

$$\Delta y_t = \gamma y_{t-1} + \sum_{j=1} \delta_j \Delta y_{t-j} + \varepsilon_t \quad (18)$$

The lags of ΔY_t "soak up" any effective framework found in the dependent variable, to ensure that not auto correlated. The test statistic just for the Augmented Dickey fuller test is described as;

$$DF_\gamma = \frac{\hat{\gamma}}{SE(\hat{\gamma})} \quad (19)$$

In every instance where the test statistics is much less favorable compared to the essential worth, the stationary alternative is dismissed rather than the null hypothesis for a device root. The test previously mentioned is valid just when ε_t is white noise.

2.3 Nonlinearity Tests

2.3.1 Brock, Dechert, and Scheinkman (BDS) Statistic

Brock, Dechert, and Scheinkman (1987) recommend an examination statistic, frequently known as the BDS test, to determine the aid presumption of any time sequence. The statistic is, therefore, totally different from some other test statistics pointed out simply because the latter generally concentrate on both the third-order or maybe second properties x_t. The main idea of the BDS test is utilizing any "correlation important" typical in chaotic time series X_t as well as observations $\{X_t\}_{t=1}^{T_k}$, determine the correlation essential as,

$$C_k(\delta) = \lim_{T_k \to \infty} \frac{2}{T_k(T_k-1)} \sum_{i<j} I_\delta(X_i, X_j) \quad (20)$$

Where $I_\delta(u,v)$ is an indicator variable which equals one in case if a $\|(u-v)\| < \delta$, and 0 else, where $\|\cdot\|$ is the sup norm. The correlation integral measures the proportion of data pairs $\{x_t\}$ within a distance of δ one another.

Define

$$C_\ell(\delta, T) = \frac{2}{T_k(T_k-1)} \sum_{i<j} I_\delta(X_i^*, X_j^*), \ell = 1, k, \quad (21)$$

Where $T_\ell = T - \ell + 1$ and $X_i^* = x_i$ if $\ell = 1$ $X_i^* = X_i^k$ and if $\ell = k$. Under the null hypothesis that $\{X_t\}$ is iid with a non-degenerated distribution function F (.), Brock, Dechert, and Scheinkman (1987) show that

$C_k(\delta, T) \to [C_1(\delta)]^k$ With probability 1, as $T \to \infty$ for any fixed k and δ. Furthermore, the statistic $\sqrt{T}\{C_k(\delta,T) - [C_1(\delta,T)]^k\}$ is asymptotically distributed as normal with mean zero and variance

$$\sigma_k^2(\delta) = 4\left(N^k + 2\sum_{j=1}^{k-1} N^{k-j} C^{2j} + (k-1)^2 C^{2k} - k^2 N C^{2k-2}\right) \quad (22)$$

Where $C = \int [F(z+\delta) - F(z-\delta)]^2 dF(z)$.

 https://doi.org/10.59568/JASIC-2024-5-1-09

Note that $C_1(\delta,T)$ is a consistent estimate of C, and N can be consistently estimated by

$$N(\delta,T)=\frac{6}{T_k(T_k-1)(T_k-2)}\sum_{t<s<u}I_\delta(X_t,X_s)I_\delta(X_s,X_u). \quad (23)$$

The BDS test statistic is then defined as

$$D_k(\delta,T)=\sqrt{T}\{C_k(\delta,T)-[C_1(\delta,T)]^k\}/\sigma_k(\delta,T), \quad (24)$$

Where $\sigma_k(\delta,T)$ is obtained from $\sigma_k(\delta)$ when C and N are replaced by $C_1(\delta,T)$ and $N(\delta,T)$, respectively. This test statistic has a standard normal limiting distribution.

3. RESULTS

3.1 Descriptive Statistics and Stationary of the Data Sets

3.1.1 Descriptive Statistics

Table 1: Descriptive Statistics of Actual Data

Parameter	Variable	
	RH at 0600	RH at 1200
Count	60	60
Mean	80.21667	61.44
StD	6.950361	11.52051
Min	60.1	35.1
25%	77.95	56.725
50%	82.15	62.9
75%	84.825	67.575
Max	93.2	83.5

Table 1 above indicates that for the actual data, there were 60 observations each for relative humidity (RH) measurements at 06:00 and 12:00, with mean RH values of 80.22% and 61.44% respectively, exhibiting variability as shown by standard deviations of 6.95% and 11.52%, and a range from a minimum of 60.1% to a maximum of 93.2% for RH at 06:00 and from 35.1% to 83.5% for RH at 12:00, with respective quartiles indicating the spread of the data around the median values.

3.1.2 Stationary Test

Table 2: ADF Unit Root Test of the Relative Humidity (R.H 06:00)

Null hypothesis:	Data are non-stationary
Alternative hypothesis:	Data are stationary

Test Statistic	P-Value	Recommendation
-3.57965	0.006	Test statistic <= critical value of -2.91284. Significance level = 0.05 Reject null hypothesis. Data appears to be stationary, not supporting differencing.

The information in Table 2 above shows that the actual series is stationary.

Figure 1 below shows that the actual values for relative humidity at 0600 is stationary without differencing.

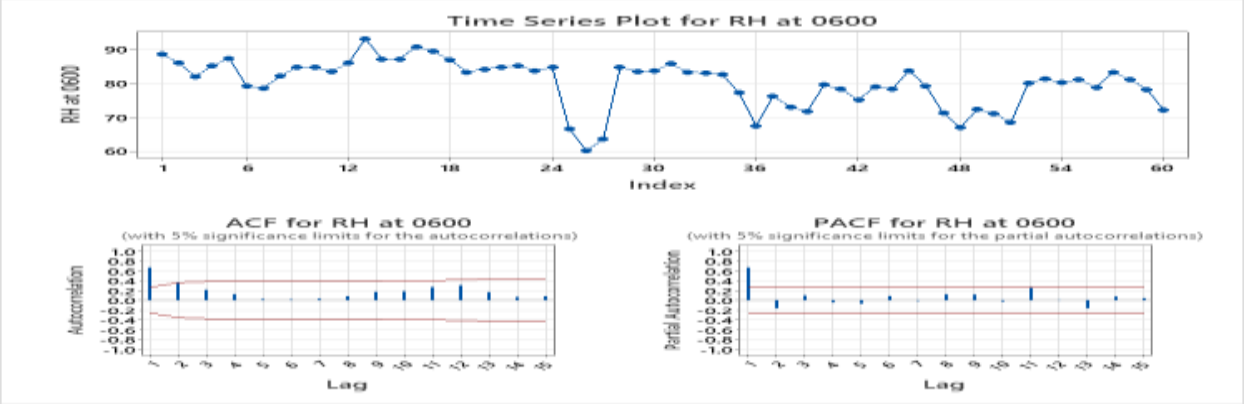

Figure 1: ACF and PACF Plots for RH at 0600

 https://doi.org/10.59568/JASIC-2024-5-1-09

Table 3: ADF Unit Root Test of the Relative Humidity (R.H 12:00)

Null hypothesis:	Data are non-stationary
Alternative hypothesis:	Data are stationary

Test Statistic	P-Value	Recommendation
-2.76197	0.064	Test statistic > critical value of -2.91194.
		Significance level = 0.05
		Fail to reject null hypothesis.
		Consider differencing to make data stationary.

The information in Table 3 above shows that the actual series is not stationary; hence first differencing method was applied to make data stationary.

Figure 2 below shows that the data for relative humidity at 1200 is stationary after first differencing.

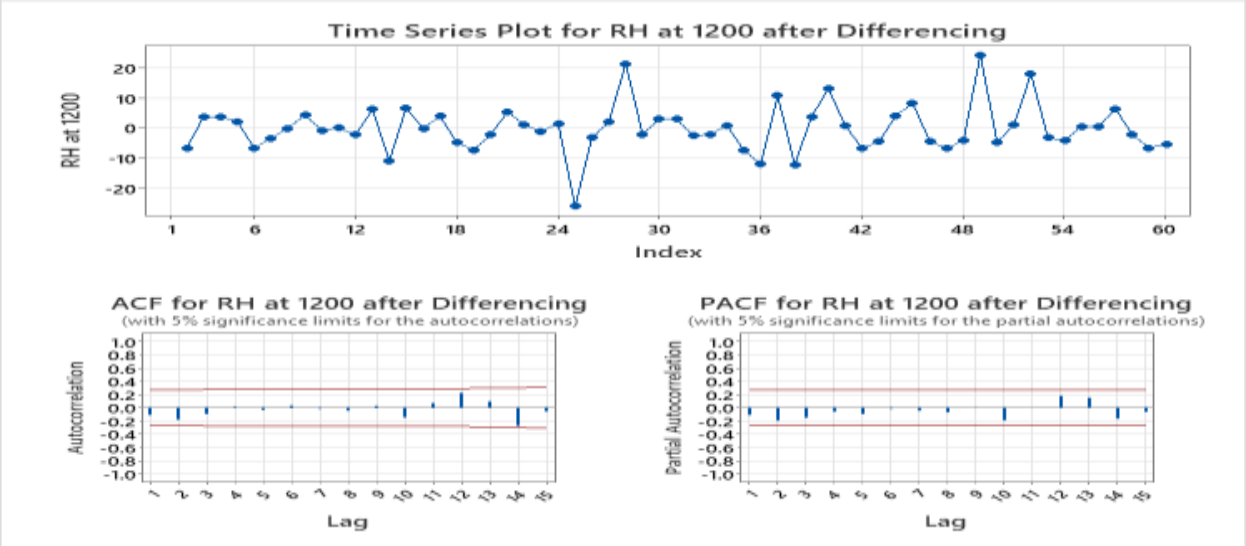

Figure 2: ACF and PACF Plots for RH at 1200 after first differencing

3.1.3 Brock, Dechert, and Scheinkman (BDS) Test for Nonlinearity of the Data

H0: The series is linear

H1: The series is nonlinear

Table 4: BDS **Nonlinearity** *Test of the Relative Humidity*

Embedding Dimension	RH (06:00)		RH (12:00)		Decision
	Z	P	Z	P	
2	7.539	0.000	9.206	0.000	Nonlinear
3	7.446	0.000	9.129	0.000	Nonlinear
4	7.479	0.000	8.990	0.000	Nonlinear
5	7.165	0.000	9.474	0.000	Nonlinear
Distance Criterion	10		17		
1st-order Correlation	0.700		0.701		

 https://doi.org/10.59568/JASIC-2024-5-1-09

Table 4 shows the BDS test results for the relative humidity (RH) measurements taken at both 06:00 and 12:00 indicate strong rejection of the null hypothesis of linearity or independence across different test orders 2 to 5, with asymptotic p-values close to zero, suggesting significant evidence of nonlinearity in the data, corroborated by a distance criterion based on first-order correlation with a value of 0.700 and 0.701 respectively. Since the series are nonlinear, next we move straight to fit a nonlinear support vector machine (SVM) model.

3.2 Fitted Nonlinear Support Vector Machine (SVM)

The categorization of relative humidity data using the nonlinear SVM method is done with radial kernels that have parameter values of C = 1.0 and σ = 0.2. The training and testing sets of data utilized in this investigation were divided into equal segments (50:50). As many as thirty training data and as many as thirty testing data were acquired. The accuracy results are displayed as follows in Tables 5 and 6:

Table 5: Testing and Training Result of the Relative Humidity

RH at 06:00						
Data	**Town**	**Precision**	**Recall**	**F1-score**	**Support**	**Accuracy**
Testing	Arua	67%	50%	57%	4	33%
	Bulambuli	0%	0%	0%	2	
	Entebbe	0%	0%	0%	4	
	Kampala	33%	100%	50%	1	
	Tororo	20%	100%	33%	1	
Training	Arua	71%	62%	67%	8	58%
	Bulambuli	55%	60%	57%	10	
	Entebbe	0%	0%	0%	8	
	Kampala	53%	91%	67%	11	
	Tororo	64%	64%	64%	11	
RH at 12:00						
Testing	Arua	67%	50%	57%	4	33%
	Bulambuli	0%	0%	0%	2	
	Entebbe	0%	0%	0%	4	
	Kampala	33%	100%	50%	1	
	Tororo	20%	100%	33%	1	
Training	Arua	71%	62%	67%	8	58%
	Bulambuli	55%	60%	57%	10	
	Entebbe	0%	0%	0%	8	
	Kampala	53%	91%	67%	11	
	Tororo	64%	64%	64%	11	

Table 5 indicates the performance of a machine learning model (SVM) evaluated on different towns for both training and testing relative humidity data sets at two different times (06:00 and 12:00). The model generally shows varying levels of accuracy (ranging from 0% to 100%) across different towns, with fluctuations in precision, recall, and F1-score metrics between the training and testing phases, suggesting potential variations in model generalization. Additionally, we noticed that the testing data's accuracy value is lower than the training data's accuracy value for both 06:00 and 12:00.

3.3 Confusion Matrices for the *Testing and Training Data*

Confusion Matrix for Training Data: **Confusion Matrix for Testing Data:**

 https://doi.org/10.59568/JASIC-2024-5-1-09

8	2	0	0	0
4	5	1	0	0
0	4	5	2	0
0	1	2	6	0
0	0	0	0	8

2	0	0	0	0
0	2	0	0	0
0	1	0	0	0
0	0	0	3	0
0	0	0	0	4

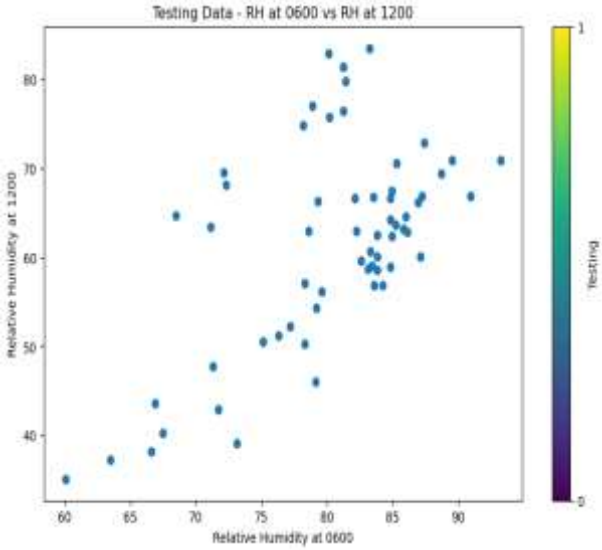

Figure 3: Scatter Plot for Testing R.H

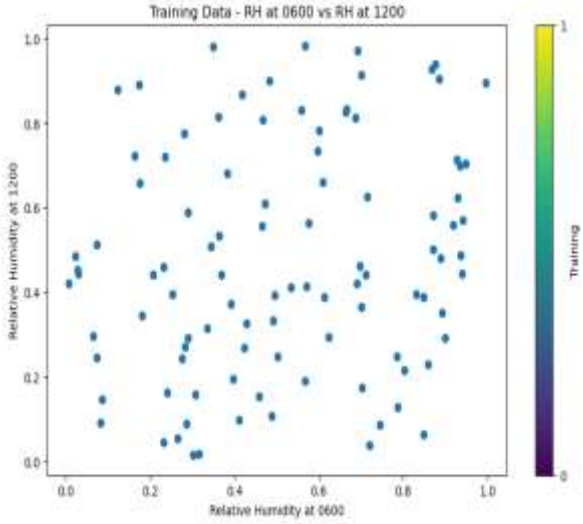

Figure 4: Scatter Plot for Training R.H

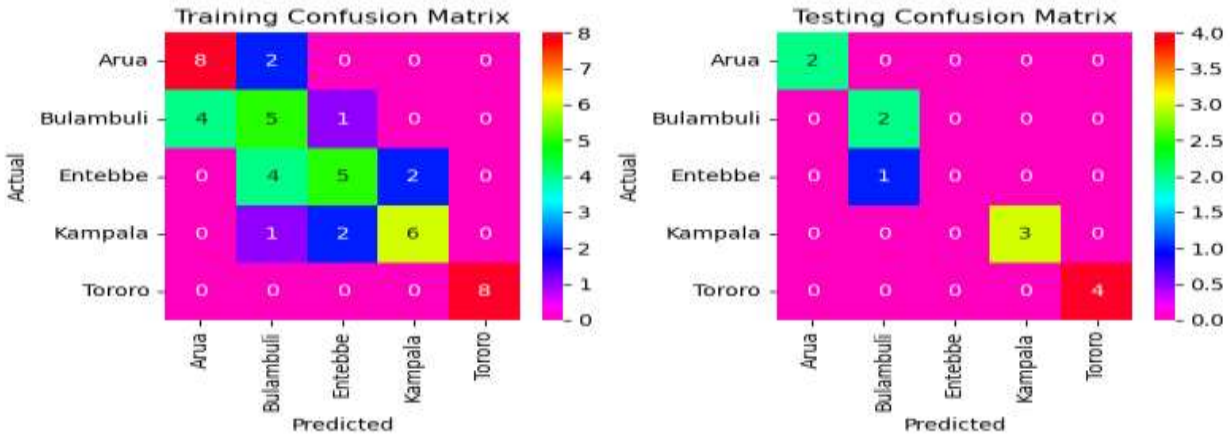

Figure 5: Confusion Matrix Plot for Training and Testing Data

4. DISCUSSION

The accuracy of the category version for the assessment data on relative humidity (RH) Measured by 06:00 is 33%. Town with greatest RH at 06:00 is Aura, having a precision of 67%, indicating that when the model predicts Arua, it is right 67% of the time. The accuracy of the classification model for the training data of relative humidity (RH) at 06:00 is 58%. Kampala in this case is the town having the highest RH of all the cities at 06:00. The model predicted Kampala to be leading with a precision of 53% of the number of time folded.

Likewise, the accuracy of the classification model for the testing data of relative humidity (RH) at 12:00 was 33%. Kampala was again identified as the town with

 https://doi.org/10.59568/JASIC-2024-5-1-09

highest RH at twelve mean averages out of all the towns: The model predicts Kampala leading with a precision of 33% of the number of time folded. The classification model's performance on relative humidity (RH) training data at 12:00 is 58 %. The model had various levels of precision as well as recall in various towns, with the highest precision and recall scoring identified in Arua and Tororo towns, indicating relative better performance in these towns when compared with other cities.

The confusion matrix for testing data shows that the classifier was 100% accurate in categorizing the indicators into particular town categories, as demonstrated by the non-zero values around the diagonal. Nevertheless, several misclassifications were noticed, particularly in the second and third category (i.e. Bulambuli and Entebbe), indicating potential issues in extending the classifier's abilities to unobserved data. Also, the training data suggested that the classifier attained varying degrees of accuracy across various town categories, with the greatest amount of correct predictions in the 5th category (Tororo), suggesting fairly correct classification performance overall. Finally, we also found that the training data for both 06 00 and 12:00 had an accuracy value of 58%. The testing data's accuracy value is (33%), indicating a 25% increase in accuracy for the training data compared to the testing data.

5. CONCLUSION

This work looks at the gap in research regarding powerful predictive models for relative humidity (RH) prediction structured to Uganda's climatic conditions, using Support Vector Machine (SVM) algorithms. The research seeks to accurately predict RH rates in Uganda using SVM, offering valuable information for decision making in areas like agriculture, urban planning and health. The research covers five selected geographical regions within Uganda, utilizing 3 years of monthly RH data sourced from the Uganda National Meteorological Authority (UNMA). The results revealed that RH data exhibits significant nonlinearity and the SVM model performed in different ways in the 5 towns and at different times, with training data displaying greater accuracy than testing data (33% vs. 58%). Generally, Arua and Kampala recorded the highest RH possessing the highest recall and precision scores, indicating somewhat better overall performance in predicting these towns when compared with others. The SVM technique for classification of relative humidity data has outstanding overall performance for the training data. Implications include better understanding of local humidity dynamics, helping informed decision making, though challenges in generalizing model performance to unseen data were observed. Future research might concentrate on refining the model to enhance generalization and investigating additional factors influencing RH dynamics to further enhance predictive accuracy and applicability in Uganda's context

6. REFERENCES

[1] Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). Classification and regression trees. *CRC Press*.

[2] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, *20*(3), 273-297.

[3] Arquad, M. A. H., & Bose, I. (2012). Decision support system, 53, 226-233.

[4] Fix, E., & Hodges Jr, J. L. (1951). Discriminatory analysis. Nonparametric discrimination: Consistency properties. *Technical report*, 4.

[5] Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1-67.

[6]Ghosh, A., & Choudhury, P. (2020). A comprehensive review on support vector machine in weather forecasting. *Journal of Big Data*, 7(1), 1-34.

[7] Gupta, A., & Sharma, S. (2018). Exploring the potential of Support Vector Machines for predicting relative humidity rates in urban areas of India. *International Journal of Environmental Research and Public Health*, *15*(10), 2210.

[8] Han, J., & Kamber, M. (2001). *Data mining concepts & techniques*. Academic Press.

[9] Kim, Y., & Park, S. (2020). Performance evaluation of Support Vector Machines for predicting relative humidity rates in rice-growing regions of South Korea. *Computers and Electronics in Agriculture*, *177*, 105717

[10] Kumar, M., Gromiha, M. Michael, & Raghava, G. P. S. (2008). support vector machine (SVM) for problem of classification in the resolution of two-class classification. *Proteins*, *71*(1), 189-194.

[11] Luo, Q., & Yang, H. (2017). Support vector machine-based ensemble model for forecasting relative humidity in the Tianshan Mountains, China.

 https://doi.org/10.59568/JASIC-2024-5-1-09

Theoretical and Applied Climatology, 130(1-2), 569-580.

[12] McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*(4), 115-133.

[13] Rahman, F., & Purnami, W. S. (2015). The principles of SVM function to decide the optimum hyperplane separating two groups. *Jurnal Sains Dan Seni ITS*, 1(1). ISSN 2301928X.

[14] Sain, H., Kuswanto, H., Purnami, S. W. & Rahayu, S. P. (2021). Classification of rainfall data using support vector machine. *Journal of Physics: Conference Series, 1763*, 1-7.

[15] Shrestha, A., & Bajracharya, B. (2019). A review on support vector machine for weather prediction. *Journal of Hydrology and Meteorology*, 13(1), 1-10.

[16] Smith, J., & Johnson, R. (2018). Investigating the effectiveness of Support Vector Machines in predicting relative humidity rates in Southeast Asia. *International Journal of Climatology, 38*(5), 1101-1115.

[17] Vapnik, V. (1995). *The nature of statistical learning theory*. Springer-Verlag.

 https://doi.org/10.59568/JASIC-2024-5-1-09