UNIVERSITY OF NAIROBI

DEPARTMENT OF PHYSICS

PHOTOVOLTAIC (PV) SYSTEM PERFORMANCE FORECASTING AND MODELLING USING REAL -TIME OBSERVATION AND WEATHER DATA

BY

RITA MWENDE

I56/20130/2019

A Thesis Submitted for Examination in Partial Fulfillment of the Requirements of the Award of Master of Science Degree in Physics of the University of Nairobi.

2021

DECLARATION

I hereby declare that this thesis is my original work and has not been submitted for research at any other university. Where other people's work or my own work has been used, it has been properly acknowledged and referenced in accordance with the University of Nairobi's requirements.

Signature: .... Date: **24/11/2021**

Rita Mwende

I56/20130/2019

Department of Physics

University of Nairobi

This thesis has been submitted with our approval as supervisors:

	Signature	**Date**
1. Dr Sebastian M. Waita Department of Physics University of Nairobi Email: swaita@uonbi.ac.ke		25-11-2021 .
2. Dr Geoffrey O. Okeng'o Department of Physics University of Nairobi gokengo@uonbi.ac.ke	...……………………	………25.11.2021………

DEDICATION

This research thesis is dedicated to my dear mother Mercy Nduku Kilonzo and Grandmother Elizabeth Mwanzia for their prayers and support.

TABLE OF CONTENTS

ACKNOWLEGMENTS

I wish to thank God for enabling me to carry out this research to completion.

I acknowledge my supervisor's Dr Sebastian M. Waita and Dr Geoffrey O. Okeng'o for their continuous support and always providing me with necessary assistance whenever I needed it. I also wish to thank the department of Physics for their financial support in order to undertake this program.

My special thanks to Mr. Boniface Muthoka, Mr. Dickson Omucheni and Mr. Arnold Bett from the Department for help with the equipment used in this study and for their technical support. I also wish like to appreciate Mr. Charles Ndung'u Ndegwa for sharing his expertise in the data science field and for the endless consultations that were a significant contribution to the completion of this project

Last but not least, I wish to appreciate my mother Mercy Nduku Kilonzo and Grandmother Elizabeth Mwanzia for their moral support and their continuous motivation and prayers. Not forgetting Mr. Bernard Kiragu for his support and encouragement.

ABSTRACT

Photovoltaic (PV) systems are an indispensable source of renewable energy supply for both commercial and domestic use in many developing countries including Kenya. However, it remains difficult to fully integrate solar energy into the power grid. This is because solar energy is intermittent and highly dependent on weather conditions. Therefore, proper modelling and assessment of the influence of environmental parameters on the PV system performance is essential. In this study, a detailed performance analysis of a 1.5 kW PV system was done to study the effect of selected weather parameters on the power output. A weather station was setup on the site to provide real-time measurements of the ambient temperature and relative humidity. Solar irradiance was measured using a HT304N reference cell and the PV module temperature measured using a HT instrument PT300N temperature sensor. A current - voltage values of the solar PV system were obtained using a current-voltage solar (I-V) analyzer. Data collection was done daily between 10:00 a.m. to 3:00 p.m. EAT at 30 minutes' interval for a period of 21 days. Data analysis and visualization was performed using the R –software statistical package and Origin 9.1 software. An interactive application based on the single diode model was also developed and the results compared to measured data. The results obtained show that the ambient temperature increases with increasing solar irradiance with correlation coefficient (P) of 0.53 and Adj R^2 of 0.27 showing a weak relationship. It was also noted that relative humidity varies inversely with solar irradiance with a correlation coefficient P of -0.50 and Adj R^2 of 0.27. Relative humidity and ambient temperature exhibited a strong negative relationship yielding a correlation of -0.94 and Adj R^2 of 0.90. It was also noted that the ambient temperature had a relatively strong linear relationship with module temperature having a correlation P of 0.84 and Adj R^2 of 0.71. The study further showed that maximum PV power output varies linearly with solar irradiance with strong positive relationship evidenced by a correlation P of 0.99 and Adj R^2 of 0.98. However, the PV system efficiency was noted to decrease with increasing solar irradiance with a negative correlation P of-0.85 and Adj R^2 of 0.72. Series resistance was found to have a strong negative non-linear relationship with solar irradiance with Adj R^2 of 1 while shunt resistance decreased non-linearly with solar irradiance of Adj R^2 of 0.64. The open circuit voltage was found to vary inversely with the module temperature with correlation P of -0.50 and low Adj R^2 of 0.25 indicating a weak relationship. The maximum power and module temperature exhibited a positive linear relationship with P of 0.70 and Adj R^2 of 0.49. It was established that the module temperature decreased with the efficiency of the PV system with P of -0.87 and Adj R^2 of 0.76. Due to the high correlation between ambient temperature, solar irradiance, relative humidity, module temperature principal component analysis (PCA) was done to remove redundant information. Support Vector regression (SVR) and random forest regression (RFR) models were therefore trained, tested and validated using data obtained from PCA to forecast real-time PV power output. SVR model employing leave one out cross validation technique (LOOCV) yielded the best model compared to *k-fold* and CV (Random resampling) cross validation techniques with root mean square (RMSE) of 40.4, Adj R^2 of 0.98 and mean absolute error (MAE) of 29.01 on training dataset and RMSE of 45.10, Adj R^2. of 0.97 and MAE of 29.27 on testing dataset. RFR model employing LOOCV yielded best model *k-fold* and CV (Random resampling) cross validation techniques with RMSE of 65, Adj R^2 of 0.95 and MAE of 51.8 on training dataset whereas for testing set RMSE of 94, Adj R^2 of 0.87, MAE of 68 were obtained. The trained models were further evaluated using validation dataset, SVR model outperformed RFR with RMSE of 43.16, Adj R^2 of 0.97 and MAE of 32.57 compared to RMSE of 86, Adj R ^{2} of 0.90 and MAE of 69 obtained from RFR model. Furthermore, an app for carrying out real-time 1.5 kW PV power output prediction based on the SVR model was developed in this study. This research work therefore demonstrates that variability of solar irradiance, ambient temperature and relative humidity have significant effect on the performance of solar PV systems and must be considered when predicting PV power output. This is a significant step towards realizing a site-specific and dynamic solar PV performance analysis and forecasting technique.

LIST OF ABBREVIATIONS AND SYMBOLS

ARIMA	Autoregressive Integrated Moving Average
ANN	Artificial neural networks
ambtemp	Ambient Temperature
Adj R^2	Adjusted R Squared
AM	Air Mass
app	Application
B	Zenith angle
CV	Cross Validation
DC	Direct current
DI	Diffuse Horizontal Irradiance
DN	Direct Normal Irradiance
E_g	Energy band gap
EAT	East Africa Time
FF	Fill Factor
FiTs	Feed-In Tariffs
G	Irradiance at STC
G_a	Measured irradiance
GE	Mean-Squared Generalization Error
GI	Global Horizontal Irradiance
Hum	Relative Humidity
I_m	Maximum current
I_{sc}	Short-circuit current
IEA	International Energy Agency
IES	International Educational Services (UK)

I_v	Output current
I_d	Diode current
I_{ds}	Dark Saturation Current
I_{sa}	Saturation current
I_p	Photo generated current
I_{Rp}	Current through the parallel resistor
I_v	Output current
I-V	Current-Voltage
k	Boltzmann Constant
KNN	*k* Nearest Neighbor
k_i	Temperature Coefficient short circuit current
k_v	Temperature Coefficient Open circuit current
KRR	Kernel ridge regression
LLAR	Linear regression-time series model
modtemp	Module Temperature
MIN	Minimize
MAE	Mean Absolute Error
ML	Machine Learning
MW	Megawatt
MoE	Ministry of Energy
n	Ideality factor
$\boldsymbol{\eta}$	Efficiency
N_s	Number of cells connected in series
NWP	Numerical Weather Prediction

NGO’s	Non- Governmental Organizations
P	Pearson’s correlation Coefficient
PC1	First Principal Component
PC2	Second Principal Component
PC’s	Principal Components
PCA	Principal Component Analysis
PV	Photovoltaic
P-V	Power-Voltage
P_i	Incident Power
P_m	Maximum power
q	Charge of an electron
R_s	Series resistance
R_p	Shunt resistance
R^2	Coefficient of determination
RMSE	Root Mean Square
RBNN	Radial basis neural network
RF	Random Forest
RFR	Random Forest Regression
RES	Renewable Energy Sources
SVM	Support Vector Machines
SVR	Support Vector Regression
STC	Standard testing conditions
T	Temperature at STC
T_a	Ambient Temperature
T_m	Module Temperature
UI	User Interface

UK	United Kingdom
V_t	Thermal Voltage at STC
V_{tn}	Thermal Voltage at module temperature
V_{oc}	Open-circuit voltage
V_m	Maximum Voltage
$\sigma_{a,b}$	Standard Deviation

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

1.1: Background of the study

Availability of sufficient, affordable and reliable energy is crucial for the wholesome development of any nation. Due to the ever-increasing world population together with advances in global technology with high power requirements, the world's energy consumption is anticipated to rise by over 50% by the year 2050 (IEA, 2019).

To date, fossil fuels are the world's primary source of energy contributing about 85% of the world's energy budget (Lenzmann and Carol, 2016). Fossil fuels, therefore play a key role in the world economy and industrial development. However, they are non-renewable, unsustainable and their combustion has detrimental effects to the environment by contributing to a rise in atmospheric greenhouse gases, destruction of ecosystems, change in weather patterns, rising sea level and melting of glaciers (IES, 2019). It is for this reason, the United Nations has called for adoption of sustainable and renewable sources of energy.

Kenya's economic growth has led to a rise in the demand for electricity from 1802 MW in 2018 to 1912MW in 2019. Demand for energy has been rising steadily by 3.6% annually (Africa Energy Series, 2020). 74.5% of Kenya's energy demand is provided by wind, hydropower, solar and geothermal power which are all renewable energy sources with fossil fuel only supplying 25.5% to the energy mix. Majority of the power is derived from hydropower supplying approximately 677MW followed by geothermal 670 MW of the total 2.7GW installed capacity (Africa Energy Series, 2020). Hydropower capacity is adversely affected by long periods of drought which have been experienced since 2015 (Africa Energy Series, 2020). Geothermal power has great potential of providing up to 10 GW power (Achieng *et al.*, 2012). However, rising investment charges, land disputes, lack trained personnel, huge grid infrastructural investment hinder its full exploitation (Samoita *et al.*, 2020).

The focus of renewable energy has shifted to solar energy due to its abundance and availability. There are two types of solar technology in use today namely photovoltaics and thermal collectors. Photovoltaics are highly popular source of renewable energy especially off grid areas. Moreover, they require low maintenance, require short construction time and are pollution free energy (Goswami, 2017).

1.1.1: Solar Energy in Kenya

Kenya receives a lot of solar radiation due to its equatorial location. Kenya experiences an average of 5-7 hours of peak sunshine with average daily insolation of between 4-6 kW/m^2. More specifically, the Northern parts and along the Lake Victoria basin generally receive higher and more consistent solar irradiance (Mark *et al.*, 2009). Solar irradiation drops to less than 3.5 kWh/m^2/day in populous regions near Nairobi, Mt Kenya, and the Aberdares, between the months of May and August (Mark *et al.*, 2009).

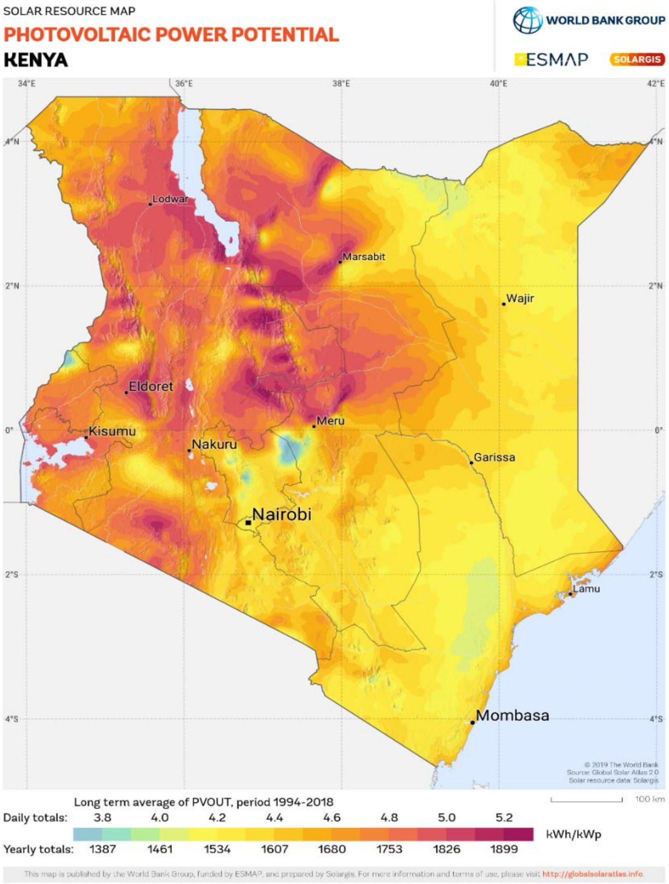

SOLAR RESOURCE MAP
PHOTOVOLTAIC POWER POTENTIAL
KENYA
WORLD BANK GROUP
ESMAP
SOLARGIS
34°E
36°E
38°E
40°E
42°E
4°N
2°N
0°
2°S
4°S
Lodwar
Marsabit
Wajir
Eldoret
Meru
Kisumu
Nakuru
Garissa
Nairobi
Lamu
Mombasa
© 2019 The World Bank
Source: Global Solar Atlas 2.0
Solar resource data: Solargis
100 km
Long term average of PVOUT, period 1994-2018
Daily totals: 3.8 4.0 4.2 4.4 4.6 4.8 5.0 5.2
kWh/kWp
Yearly totals: 1387 1461 1534 1607 1680 1753 1826 1899
This map is published by the World Bank Group, funded by ESMAP, and prepared by Solargis. For more information and terms of use, please visit http://globalsolaratlas.info

Figure 1.1: Map showing the PV potential power generated from 1994-2018 in Kenya (Solargis, 2019).

Kenya’s solar market is among the most well-established in Africa with its roots extending to as early as the 1970’s with PV sales estimated at more than 1.2 MW market per year (Mark *et al.*, 2009). The PV market has grown steadily for the past ten years at over 10 % annually. Kenya's total installed PV capacity is 100MW. This capacity is projected to grow yearly by 15% (Samoita *et al.*, 2020).

The solar energy market is composed of tourism , telecom off-grid (community and household) and small-scale business electrification. In the late 1980’s, off-grid household electrification kicked off in the most densely inhabited rural areas with television signals. In the early to mid-1990’s coffee and tea “boom period” small-scale farmers from remote areas started purchasing residential solar electric systems. Off-grid household market is highly competitive and mature, however concerns of sub-standard quality of components and poor installation practices have recently risen (Mark *et al.*,2009). Off-grid community/ institutional systems include many active NGO’s, schools, churches missions, hospitals and other amenities that provide services to remote areas e.g. West Pokot also use solar power to operate their project. The tourism market includes game parks, hotels that are off-grid in the tourism industry while the telecom market is primarily made up of the mobile networks that use solar powered base stations.

The Kenyan government has also implemented policies such as Feed-In tariffs and 2006 Power Act No.12 that encourages the utilization of renewable energy sources (RES) to improve electricity generation in the country (Ministry of Energy, 2012). Feed-in tariffs (FiTs) are a policy that forces power providers to sell electricity generated by renewables at a predetermined rate. These government efforts have led to increase in PV system installations all over the country such as the largest solar farm of 55 MW in Garissa.

Solar power generation by PV depends heavily on weather variability. Due to the fluctuating weather conditions the question of reliability of solar power and its ability to satisfy the demand for energy remains an important unsolved problem in solar energy research. In this study a detailed analysis of the effect of relative humidity, solar irradiance ,ambient temperature and module temperature on the performance of a 1.5 kW solar PV system was performed. Machine learning- based predictive models to predict PV power output under varying weather conditions were built.

1.2: Statement of the Problem

Solar PV systems have a huge potential for generating vast amounts of electric power, but their performance and power output is often highly variable, and heavily dependent on fluctuations in solar irradiance and other weather conditions. The inherent fluctuating nature of solar power sources poses a major challenge in the quest to fully integrate solar energy power plants into existing power grids without compromising on the stability of the power output. Therefore, as the number of solar PV systems and solar grid connected solar power plants installations increase, there is an urgent need to carry out research aimed at developing techniques and models with the capability of performing accurate real-time site-specific performance analysis and power output forecasting.

This research work aimed at carrying out detailed analysis and modelling of the effect of fluctuating weather parameters and solar irradiance on the performance of a 1.5 kW solar PV system, as well as develop a machine learning- based technique for performing real-time solar PV power system performance under varying weather conditions.

1.3: Objectives

1.3.1: Main Objective

To perform detailed analysis and modelling of PV solar systems performance using real-time observations and weather data as well as building a flexible and adaptable solar PV power forecasting model employing machine learning techniques.

1.3.2: Specific Objectives

The specific objectives of this study were:

1. To obtain real time measurements of solar irradiance, relative humidity, module temperature and ambient temperature for analysis.
2. To obtain real time performance data of 1.5 kW PV solar system for analysis.
3. To model the 1.5 kW PV system performance (I-V curve and power output) and develop a forecasting model based on machine learning techniques.

1.4: Justification and Significance of the study

The fast-increasing worldwide installation and use of solar PV systems has made it necessary to carry out research aimed at developing accurate and site-specific techniques and systems able to carry out real-time PV system performance analysis and output power forecasting. Particularly, solar energy stability is highly affected by variations in solar irradiance and various weather parameters. Most solar panels are flash tested at 1000 W/m^2, 25°C Air mass 1.5 which are unrealistic owing the fact that these conditions during outdoor operations nearly never occur.

Furthermore, solar energy is highly intermittent and heavily depended on weather fluctuations, proper energy budgeting and planning, requires the development of reliable predictive and forecasting models able to provide accurate performance forecasts and modelling information for PV solar systems power output. This has the advantage of improving stability in power supply by providing predictions in PV power systems generation that are crucial for system controllers and future energy planning.

Research aimed at developing accurate PV systems power analysis and prediction models is therefore crucial in helping realize value and maximize returns from investments in Kenya's solar energy sector and therefore help reduce the existing barriers to the effective contribution of solar power to the national grid as well as domestic consumption. This research work is an endeavor in that direction.

CHAPTER TWO: LITERATURE REVIEW

2.0: Chapter overview

In this chapter, PV power forecasting methods are discussed. It further gives reviews on statistical methods used in PV power forecasting and also highlights gaps in existing research.

2.1: PV Power forecasting methods

The key methods used for PV power forecasting include; physical, statistical and hybrid methods. Physical methods are implemented by predicting the position of the sun and assuming that the other relevant climatic conditions negligible and are used to predict solar irradiance which is in turn utilized to predict PV power output on assumption that solar irradiance is directly proportional to PV power output. Physical PV forecasting methods are broken down into two namely: numerical weather prediction (NWP) and satellite sky imagery.

Numerical weather prediction methods (NWP) use numerical atmospheric and oceanic mathematical models to predict solar irradiance which is assumed to directly proportional to PV power output. NWP is used to predict accurately the local solar irradiance received in an area and the cloud cover for up to 15 days (Wan *et al.*, 2015). However, NWP can only predict an average solar irradiance of a grid and never the precise value for exact coordinate which may lead to inaccurate predictions of the PV power a solar PV system may produce at a certain location (Isaksson and Conde, 2018). The operation of NWP models requires a lot of data and hence incurs high computational costs. In addition, cloud characterization by NWP can remain unresolved for hours due to its limited spatial resolution. They are also based on mathematical models that have assumptions and constraints which may pose inaccuracies in the solar irradiance predictions.

Satellite sky imagery on the other hand uses satellite images to detect cloud motion to predict global horizontal irradiance with good accuracy for up to six hours (Wan *et al.*, 2015). The clouds can either be detected using the cloud motion vectors and total sky imagers that use image processing method (Jang *et al.*, 2016). Satellite sky imagery solely predicts the PV power output according to the condition of the sky either if it’s a clear sky, overcast or with scattered clouds. It can only do short term predictions due to the intermittency of the cloud movement (Schmidt *et al.*, 2017). In addition, lack of frequent sky images updates, poor geographic registration of satellite images, lack of knowledge on cloud altitudes can lead to inaccurate PV solar power predictions (Kostylev and

Pavlovski, 2011). The estimation of the sky clearness index through computation of pixel range is also challenging. This method also assumes that cloud cover is always in a steady state which is not realistic (Kostylev and Pavlovski, 2011).

Statistical methods use past irradiance data and PV power output on notion that the future will be a repeat of the past conditions. Statistical methods are divided into two categories; persistence models (classical methods) which assume the future is same as previous value obtained and time series or machine learning models that use historical data over a certain time frame to make predictions i.e. hours, days or even months. Hybrid methods merge two forecasting techniques to enhance predictive accuracy for instance NWP can be used with Artificial Neural networks (ANN). Hybrid methods have been adopted to predict power generation data from solar PV modules.

There exists extensive literature on the prediction of the generation of photovoltaic energy such as Rahman *et al.*, (2018), Khandakar *et al.*, (2019), Aliberti *et al.*, (2018), Cervone *et al.*, (2017) and Alanazi *et al.*, (2017). Most research work is focused on either predicting solar irradiance or directly predicting the PV power generation (Zafarani *et al.*, 2018).

2.2: Advances in PV power forecasting

Statistical methods for PV power forecasting have gained popularity because they are data driven. Statistical methods are much simpler to implement, require less input data than traditional methods hence have low computational costs (Giorgi *et al.*, 2014). They also study the relationships between the weather variables and can determine variable importance (Al-Dahidi *et al.*, 2018).

The existing prediction models for solar irradiance use climate parameters like solar irradiance sunshine hours, cloud cover, temperature data retrieved from satellite observations and or ground stations (Salcedo-Sanz *et al.*, 2014). However solar radiation forecasting is mainly used on assumption that solar radiation is directly proportional to PV power output (Wu and Lou, 2011). However, this assumption has led to inaccurate forecasts, hence researchers have shifted their focus on directly predicting PV power output using solar irradiance, temperature, wind speed, cloud cover and relative humidity (Khandakar *et al.*, 2019). This approach has yielded forecasts that have a higher accuracy than solar irradiance only based prediction models (Madhiarasan and Deepa, 2017).

The choice of weather variables relevant for PV forecasting is crucial for the achievement of an accurate prediction. Bhattacharya *et al.*, (2014) investigated the impact of temperature and wind speed on efficiency of a PV module for one year. The results showed a positive linear trend between module efficiency with both ambient temperature and wind speed. Whereas Panjwani and Narejo, (2014) reported on the influence of humidity on PV power production. The results showed that when the humidity was high the power output decreased up to 15-30%. Other environmental factors influencing the efficiency of the PV module were neglected in this research. Mekhilef *et al.*, (2012) did a study on the impact of humidity, air velocity and dust on efficiency of solar module. Dust accumulation decreased the efficiency while high humidity resulted to cell encapsulate delamination. Humidity and irradiance showed a non- linear relationship; hence humidity has greater impact on the current than on voltage.

Khandakar *et al.*, (2019) developed a model to monitor, assess and quantify photovoltaic performance using different weather parameters namely wind speed, solar irradiance, ambient temperature, dust accumulation and PV temperature. Prediction models developed utilized Artificial neural network linear regression, decision tree and Gaussian process regression model, they used the data collected on site for the study. ANN prediction model output power was the best. They discovered that incorporating correlation feature selection and relief feature selection together with the ANN yielded more accurate power output predictions.

Yerrapragada *et al.*, (2013) trained models based on linear regression, locally-weighted linear regression and support vector regression using UV irradiance, maximum wind speed, temperature, barometer pressure, wind chill, humidity, dew point, timestamp, wind speed, rain, heat index, extra-terrestrial irradiance and wind direction as input data. Principal component analysis reduced the dimensionality and it improved the performance of the model. Least-square variation support vector regression performed best when used together with the auto-regression integrated moving-average (ARIMA) model with an RMSE of 40.16.

Almohri *et al.*, (2014), used time series data to forecast output power using linear regression-time series model (LLAR), Radial Basis neural network (RBNN) and Kernel ridge regression (KRR). The input data included ambient irradiance, ambient temperature, fixed array insulation and fixed array

temperature. Feature selection techniques Lasso and backward stepwise techniques were used for data preprocessing. LLAR performed best compared to the rest of techniques used.
Isaksson and Conde, (2018) benchmarked on comparing several machine learning techniques with time series models across five distinct locations in Sweden. They concluded that time series model was complex due to non-stationary nature of energy output whereas machine learning techniques were easier to implement. The time series model was also not able to capture different seasonal trends. The ANN and Grade boosting regression performed best compared to *k* Nearest Neighbor (*k*-NN) and Lasso regression. It was noted that NWP data was not for the exact location of installations hence reduced accuracy of prediction. They recommended building of different models for different seasons pointing out that using data for a year or more may generalize the model hence yield inaccurate predictions.

Kanwal, (2018) also developed a Gaussian Process Regression (GPR) and Support vector regression algorithms (SVR) to forecast PV power using temperature and irradiance in Pakistan. Input data was trained, tested, and validated. The SVR based trained model was more accurate than GPR. However, the model did not study the effect other weather parameters on PV performance. Abuella and Badrul, (2016) presented PV power forecasts for 24 hours ahead using twelve variables. The support vector regression, multiple linear regression and ANN models were created. The SVR forecasting model was reported to have performed best compared to multiple linear regression and reported to be more robust than ANN. Whereas Paulin and Praynlin, (2016) used ANN to predict hourly solar power output. The input data used for the predictive model included humidity, wind speed, solar irradiance, temperature, cloud coverage, output AC power and output DC power collected for one month.

Kim *et al.*, (2019) developed a PV power forecast model using a selection of unknown weather parameters with confirmed weather forecast data. Support vector machines, *k*-NN, adaptive boosting, random forest regression, classification and regression tree, and artificial neural network prediction models were developed. The results demonstrated that the random forest regression (RFR) achieved an R-square value of 70.5% which was the best model. However, the model was over-fitted leading to over prediction on certain days. This was attributed to the physical location of the power plant.
Data quality, proper data collection and preprocessing is essential for a good solar PV power predictive model. (Isaksson and Conde, 2018) noted that using numerical weather prediction irradiance forecasts

data as an input is not ideal since it fails to provide data for the exact coordinates the PV systems were installed. The weather forecast agencies also tend to report after a few hours which can lead to data inconsistency. Almohri *et al.*, (2014), Khandakar *et al.*, (2019), Isaksson and Conde, (2018), Yerrapragada *et al.*, (2013) noted that implementing feature selection technique improved the accuracy of machine learning model by reducing redundancy in the data.

Several concerns about training models with data acquired over large periods of time were raised because they resulted in inaccuracy in the model. Isaksson and Conde, (2018) recommended training models on data according to seasons. They noted that data collected over a year or more can cause inaccurate predictions due to changes experienced in every season. As a result, it is suggested that data obtained during a specific season be used to train a model to predict power output for the same season, which would improve the predictive model's accuracy. Hence exploration of statistical methods based on onsite seasonal data collection and implementation of features selection can be used to build flexible and adaptable short-term PV power forecast methods.

This work aimed at creating interactive models that accurately predict PV power output using real-time observations and weather data using support vector machine and random forest, and analyzed the relationship between weather variables and their influence on the performance of a solar PV system.

CHAPTER THREE: THEORETICAL FRAMEWORK

3.0: Chapter overview

This chapter discusses the theoretical background of solar radiation, solar cells and the equivalent circuit. The theory on exploratory data analysis and principal component analysis is also covered. The chapter further discusses the support vector regression and random forest regression theory.

3.1: Solar radiation

Solar radiation is radiant energy (light and heat) released by the Sun. The Sun's surface temperature is about 6000K hence it behaves like a nearly perfect blackbody. The average energy received from the Sun at an average area of 1m^2 outside the earth's atmosphere is known as the solar constant (Zeman, *et al.*, 2014). However, scattering by air molecules and water absorption attenuates solar radiation before reaches the Earth's surface.

Solar radiation that reaching the earth's surface differs in both intensity and spectral composition at various locations. Solar irradiance covers the shortest distance to reach the earth's surface when the sun is overhead. The optical air mass (AM) is the ratio of sunlight's actual travel length to its minimal path length and is computed using equation 3.1 (Zeman, *et al.*, 2014)

$$AM = \frac{1}{cos\,(B)} \tag{3.1}$$

where B is the zenith angle.

When the sun is directly overhead the AM is 1 while outside the Earth's atmosphere, we have air mass equal to zero (AM0) and air mass is 1.5 (AM1.5) on the Earth's surface corresponding to the spectrum of the solar radiation with zenith angle equal to 48.19°. AM1.5 is most widely adopted standard test conditions (STC) for solar cells at 1000 W/m^2 and temperature 25°C. An illustration of AM1, AM0 and AM1.5 Figure 3.1

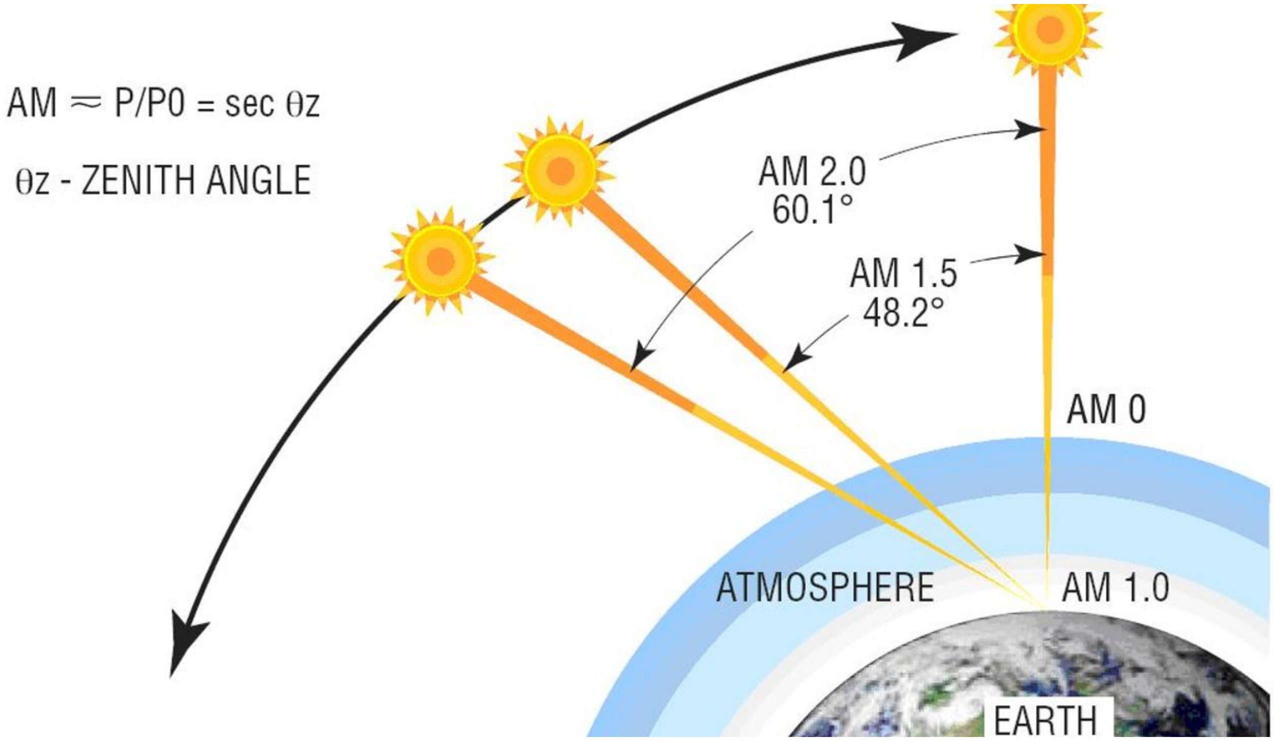


Figure 3.1: Illustration of air mass, AM0 outside the earth's atmosphere, AM1 at the earth's surface for normal incidence, AM1.5 at earth's surface at zenith of 48.2 and AM 2.0 on the earth's surface at zenith of 60.1 (Jeong, 2021).

The irradiance received by the earth when the sun is overhead (zenith) is known as the direct normal irradiance (DN). The irradiance that undergoes scattering or reflection by atmospheric components is known as the diffuse horizontal irradiance (DI). The total irradiance received on earth's surface is known as global horizontal irradiance (GI) and is expressed by equation 3.2 (Zeman, *et al.*, 2014)

$$GI = DI + DN * cos\,(B) \tag{3.2}$$

The spectral irradiance at the sun surface, outside the earth's atmosphere and at the surface of the earth is shown in Figure 3.2

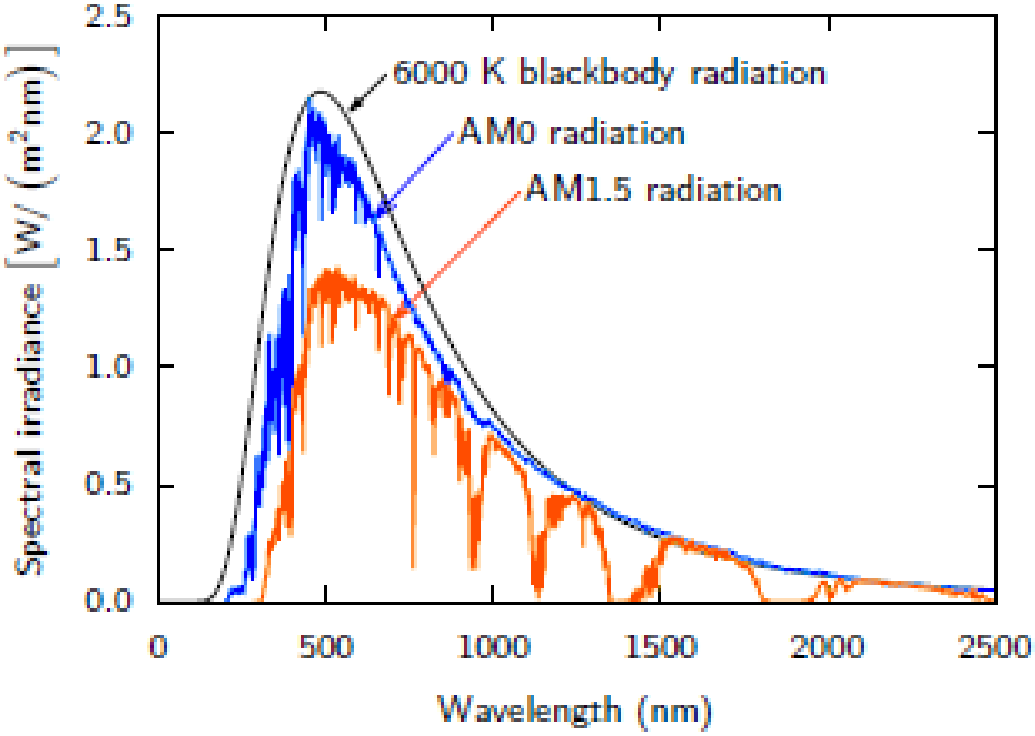


Figure 3.2: Blackbody spectrum at 6000 K, AM0 spectrum and the AM1.5 spectrum (Zeman, *et al.*, 2014).

3.2: Solar cells

Solar cell is the basic unit used in the conversion of light into electrical energy. A collection of connected solar cells is known as PV module. Illustration of a solar cell and PV module are shown in Figure 3.3. Solar cell operations are built on the photovoltaic effect which creates a potential difference at the junction of two dissimilar materials when struck by radiation. Absorption of incident photons results into the formation of electron-hole pairs when the photon energy is greater than the band gap energy (Zeman, *et al.*, 2014). Electron-hole pairs only exist for a limited time before recombination occurs. If carrier recombination occurs no power is generated. Hence the p-n junction inhibits carrier recombination by separating the electrons and the holes through the electric field. The charge carriers are obtained from electrical contact of the solar cells and transferred to the external circuit to perform work, thus generating electric energy (Zeman, *et al.*, 2014).

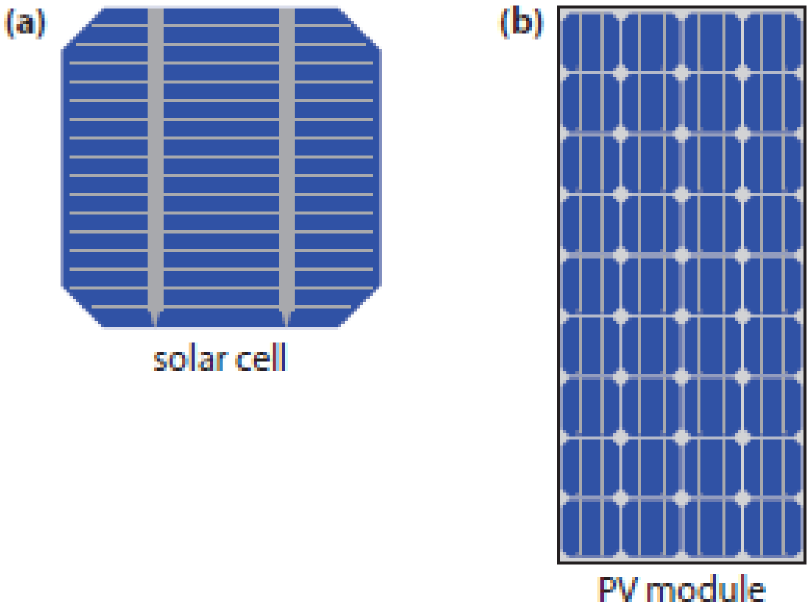


Figure 3.3: Illustration of Solar cell and PV module (a) Solar cell (b) PV module (Zeman, *et al.*, 2014).

3.2.1: External solar cell parameters

A solar cell is defined by the following key parameters: open-circuit voltage (V_{oc}), short-circuit current (I_{sc}), fill factor (FF) and maximum power (P_m). The short circuit current is the highest current that travels via the outer circuit when voltage over the solar cell is zero. The open circuit voltage (V_{oc}) refers to the maximum voltage that a solar panel generates at zero current (Zeman, *et al*., 2014).

The ratio of maximum power (P_m) of the actual PV module to the product of the open circuit voltage (V_{oc}) and short circuit current (I_{sc}) of an ideal solar cell is known as the fill factor (FF) and is given via equation 3.3 (Zeman, *et al*., 2014)

$$FF = \frac{P_m}{V_o\, I_{sc}} \tag{3.3}$$

where: $P_m = I_m\, V_m$, I_m is maximum current, V_m is maximum voltage

The fill factor determines the quality of a solar cell with a value varying from 0.7 to 0.8 for a good solar cell while 0.4 for a low-quality solar cell (Zeman, *et al*., 2014). On the other hand, the module efficiency of the PV module is the ratio of the maximum power generated to the incident power. Module efficiency (η) depends on the active area that the solar cells cover as expressed in equation

3.4. The widely used standard for quantifying efficiency is by using incident power (P_i) of 1000 W/m^2 for the AM 1.5.

$$\eta = \frac{P_m}{AP_i} \equiv \frac{I_{sc}V_{0c}FF}{AP_i} \tag{3.4}$$

where is A is area of module

Zeman, *et al.*, (2014) quoted that solar cell efficiency of crystalline silicon solar cell lies between 17% to 18%. Solar cells are especially highly susceptible to high temperatures and radiation damage. High PV module surface temperature and ambient temperatures cause overheating of PV panel hence reducing its efficiency considerably. The nominal operating cell temperature (NOCT) is temperature at an irradiance of 800W/m^2, temperature of 20°C and wind speed of 1m/s. It is applied in the determination of a PV module temperature (T_m) as shown in equation 3.5 (Zeman, *et al.*, 2014)

$$T_m = T_a + \frac{T_n - 20℃}{800} G_a \tag{3.5}$$

where; T_a is the ambient temperature, T_n is the nominal temperature and G$_a$ is the actual irradiance.

When the module temperature is high, the open circuit voltage decreases outweighing the short circuit current hence decreasing the overall system efficiency, the maximum power output and fill factor. The influence of temperature on the output of solar cells is expressed by temperature coefficients given by manufactures data sheet on the PV module (Zeman, *et al.*, 2014). One can therefore estimate the temperature of the module when the temperature coefficient is known.

3.3: The Equivalent circuit

The current-voltage (I-V) curve is essential in the characterization of PV module. The solar I-V curve is a superposition of PV module's diode I-V curves under illumination and in the dark. Figure 3.4 illustrates I-V characteristics curve.

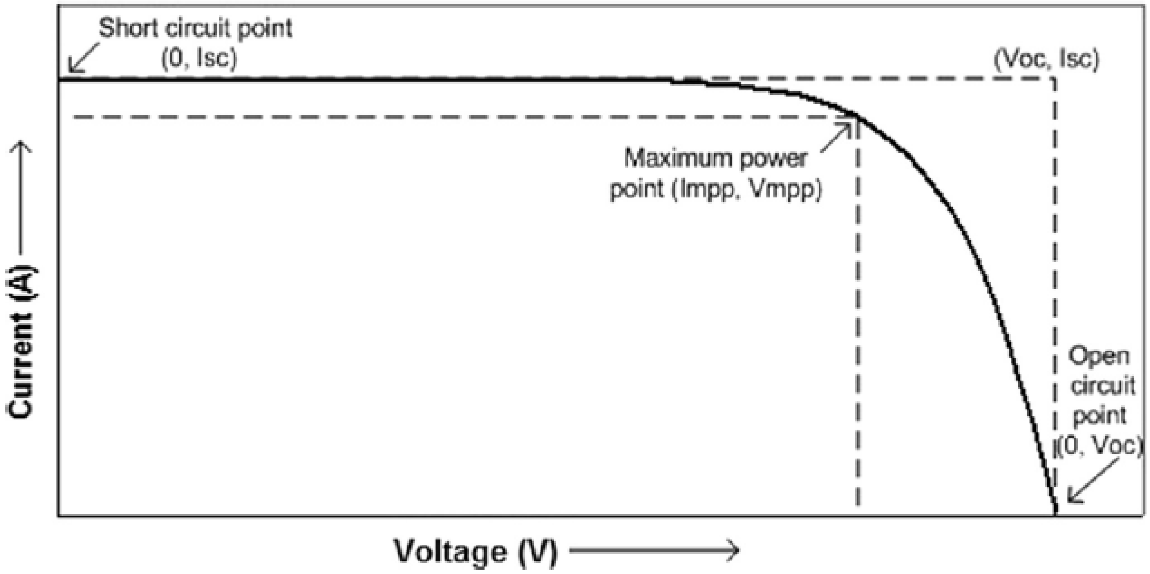


Figure 3.4: I-V characteristic curve of a PV module (Elkholy and El-ela, 2019).

The standard equivalent single diode model is utilized to model the performance of a solar cell. It is grounded on the Kirchhoff current law where the output current generated from a solar cell is equivalent to the photo current produced minus the current via the diode and the current via the parallel resistor as shown in equation 3.6 (Elkholy and El-ela, 2019)

$$I_v = I_p - I_d - I_{Rp} \tag{3.6}$$

where I_v is the output current, I_p is the photo generated current, I_d is the diode current and the I_{Rp} is current through the parallel resistor.

An illustration of the equivalent single diode model see Figure (3.5).

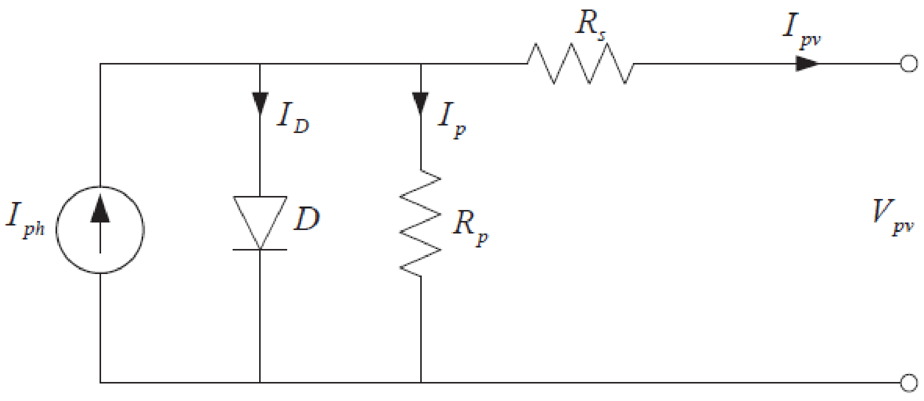


Figure 3.5: Equivalent single diode model (Elkholy and El-ela, 2019).

The diode current (I_d) is defined by equation 3.7 (Elkholy and El-ela, 2019)

$$I_d = I_{ds}\left[exp^{\frac{q\left(V_{pv}+R_sI_v\right)}{N_snkT_m}}\right] - 1 \tag{3.7}$$

where I_{ds} is the dark saturation current, q is charge of the electron, n is the ideality factor, k is Boltzmann constant, N_s is the number of cells connected in series, and R_s is series resistance.

Diode current can be written as shown in equation 3.8;

$$I_d = I_{ds}\left[exp^{\frac{\left(V_{pv}+R_sI_v\right)}{nV_{tn}}}\right] \tag{3.8}$$

where V_{tn} represents the thermal voltage at module temperature given by equation 3.9

$$V_{tn} = \frac{N_skT_m}{q} \tag{3.9}$$

The dark saturation current (I_{ds}) and its dependence on temperature is expressed via equation 3.10 (Elkholy and El-ela, 2019)

$$I_{ds} = I_{sa}\left[\frac{T_m}{T}\right]^3 exp^{\left(\frac{qE_g}{nk}\left[\frac{1}{T}-\frac{1}{T_m}\right]\right)} \tag{3.10}$$

where I_{sa} is saturation current, k is Boltzmann constant and T_m is module temperature, T is temperature at STC and E_g is energy band gap

The saturation current is given by the equation 3.11

$$I_{sa} = \frac{I_s}{exp\,(\frac{V_{oc}{'}}{nVt}) - 1} \tag{3.11}$$

where V_t is the thermal voltage at STC, V_t is equation 3.12

$$V_t = \frac{N_s kT}{q} \tag{3.12}$$

$$V_{oc}{'} = V_{oc} + k_v * (T_m - T) \tag{3.13}$$

$V_{oc}{'}$ is the open circuit voltage, V_{oc} is the open circuit voltage, k_v is the temperature coefficient at V_{oc}

The quantity of irradiance influences production of charge carriers hence influences the photo current produced by the cell (Dhar *et al.*, 2010). The photo generated current (I_p) expressed in equation 3.14 depends on irradiance and is also affected by temperature (Elkholy and El-ela, 2019).

$$I_p = \frac{G_a}{G}(I_p + k_i(T_m - T) \tag{3.14}$$

where; G is irradiance at STC, k_i is temperature coefficient for short circuit current.

$$I_{RP} = \frac{V_o{'} + R_s I_v}{R_p} \tag{3.15}$$

where R_p is shunt resistance, R_s is series resistance, I_{RP} is current through the parallel resistor.

Series resistance is obtained by finding the inverse of slope I-V at open circuit voltage V_{oc} shown equation 3.16 (Diantoro *et al.*, 2018)

$$R_s = -\frac{\Delta V_{oc}}{\Delta I_o} \tag{3.16}$$

where I_o is current at V_{oc}

Shunt resistance can be calculated from the inverse of slope I-V at short circuit current I_s as shown by equation 3.17 (Diantoro *et al.*, 2018)

$$R_p = -\frac{\Delta V_{sc}}{\Delta I_{sc}} \tag{3.17}$$

where V_{sc} is voltage at short circuit current

3.4: Exploratory data analysis (EDA)

Exploratory data analysis involves a careful and structured way of studying data to discover unexpected trends, outcomes and outliers from the dataly7 which creates a good starting point for hypothesis generation rather than only drawing statistical conclusions (Behrens, 1997). Data visualization is key in EDA, some of graphical visualizations used include the boxplots, histogram, scatter plot and many more. In earlier stages of data analysis, it is advised to plot data as measured because data preprocessing techniques such as averaging can lead to misinterpretation of data by distorting one’s visual impression of the data. However, inaccuracy is an essential feature of real measured or collected data, even under the best measuring conditions. For this reason, EDA explores the five-number summary which explores the sample minimum, lower quartile, upper quartile, median, and sample maximum which is better than classical summaries mean and standard deviation which don’t interpret the data distribution wholly (Pearsons, 2018).

Pearson correlation coefficient (p_{ab}) quantifies the linear relationship between two numerical variables and is popular because it is easy to compute and to interpret (Pearsons, 2018). For two random variables a and b, the Pearson correlation coefficient is expressed in terms of expected value ,is given by equation 3.18 (Pearsons, 2018)

$$p_{ab} = \frac{E\{(a - E\{a\})(b - E\{b\})\}}{\sigma_a \sigma_b} \qquad (3.18)$$

where the standard deviation σ_a, σ_b for variables a and b are given by equation 3.19 and 3.20:

$$\sigma_a = \sqrt{E\{(a - E\{a\})^2\}} \qquad (3.19)$$

$$\sigma_b = \sqrt{E\{(b - E\{b\})^2\}} \qquad (3.20)$$

The Pearson's correlation coefficient (p_{ab}) ranging from -1 and +1, with p_{ab} of +1 denoting perfect positive correlation , with -1 denoting a perfect negative correlation and a correlation of zero shows that the variables are statistically independent. p_{ab} relationship with linear regression is shown by equation 3.12 (Pearsons, 2018)

$$\widehat{m} = \hat{P}_{ab} \frac{\hat{\sigma}_a}{\hat{\sigma}_b} \qquad (3.21)$$

where $\widehat{m}$ is the slope .

3.4.1: Principal Component Analysis (PCA)

The Principal Component Analysis (PCA) allows the extraction of important information from data and presenting them in a set of new uncorrelated features known as principal components. The new data is presented as linear combinations of the original variables with the information in a given PC corresponding to its total variance.

According to the PCA, the directions with the biggest variations are the most significant. PC1 which is the first PC direction is along the largest variation while PC2 is the second most important with direction orthogonal to PC1 axis. PC1 is a linear combination of variable S_1,$S_{2,.}$$S_k$ as shown in equation 3.22 (Ho, 2019)

$$PC1 = a_{11}S_1 + a_{12}S_2 + \ldots a_{1k}S_k \qquad (3.22)$$

It can also be written in matrix form as

$$PC1 = a^T S$$

where a^T are the weights vectors

Since PC1 should have largest variance one can manipulate value of the weights, hence a constraint is employed where the sum of the squares of weights are considered to be unitary as shown in equation 3.23 (Ho, 2019).

$$a_{11}{}^{2} + a_{12}{}^{2} + \ldots a_{1k}{}^{2} = 1 \tag{3.23}$$

When conducting PCA the data is first centered by subtracting the average of all column means from each observation, the covariance and variance of the data is obtained and the Eigen values and Eigen vectors of the matrix are calculated. The Eigen value represents the measure of the amount of variance in each PC and are plotted in scree plot. Loadings plots are used to show how strongly the variables impacts a principal component (Pearsons, 2018). PCA enables one to identify hidden patterns in data, reduces dimensionality in data by removing redundancy in data and also helps one identify correlation between variables. It is highly recommended for dataset that exhibits high correlation which indicates redundancy in the dataset, it reduces data to PC'S which are uncorrelated and leads to overall increase in the accuracy of machine learning prediction models (Kassambara, 2011).

3.5: Machine learning techniques

Machine learning involves building mathematical models for data analysis purposes (Vanderplas, 2016). Over the years, this technique has gained popularity because of its easy adaptation to changing input data without necessarily programming.

Machine learning techniques are split into supervised, unsupervised and reinforcement learning. Supervised learning deals with modelling the link between measurable properties of data and labeled data. The model created applies known labels to unknown new data. Supervised learning is divided into classification and regression. The labels in classification are discrete categories, whereas the labels in regression are continuous values (Vanderplas, 2016). Some of the commonly used supervised learning techniques are linear regression, random forest regression and support vector machines.

Unsupervised learning involves modelling without attaching any labels to the features of a dataset. It usually includes clustering and dimensionality reduction. Clustering detects unique groups of data, whereas dimensional reduction reduces data to a more concise form. Reinforcement learning refers to the machine making sequential judgements from rewards or punishments received as a result of past acts.

3.5.1: Support vector machines

Support vector machines (SVM) is used to perform both classification and regression. It involves the construction of a separation hyperplane or collection of hyperplanes to execute regression on high dimensional data. When the algorithm gets labeled training data it forms the optimum hyperplane which separates new sample data with main the goal being to find a hyperplane $f(x)$ that has most deviation (ε) from the training data and should be leveled as possible.

Hyperplane $f(x)$ is expressed by linear equation 3.24

$$f(x) = w_i x_i + b \tag{3.24}$$

where b is the slack variable

In SVR, the set absolute error or deviation from the hyperplane should be less or equal to the specified margin called the maximum error ε whose value parameter can be tuned to achieve high accuracy in a model. To ensure the flatness of the hyperplane the w is made as small has possible by optimizing the problem to give equation 3.25 (Smola and Olkopf, 2004):

$$Min \frac{1}{2} |w|^2 \tag{3.25}$$

subject to equation 3.26

$$y_{i-} w_i x_i + b \leq \varepsilon \tag{3.26}$$

$$w_i x_i + b - y_i \leq \varepsilon$$

where ε is the deviation, Min is minimize.

Most case errors may occur beyond the ε we denote the deviation from the margin as ξ_i, Equation (3.25) now expressed as shown equation 3.27 (Smola and Olkopf, 2004).

$$Min \frac{1}{2} |w|^2 + C \sum_{i=1}^{l} \quad \xi_i \tag{3.27}$$

Constraints are expressed in equation 3.28

$$y_i - w x_i - b \leq \varepsilon + \xi_i \tag{3.28}$$

$$w_i x_i + b - y_i \leq \varepsilon + \ \xi_i$$

where $C \ > 0$ is the penalty parameter of the error term.

When C increases the tolerance for points outside the ε also increases and as C approaches zero the tolerance approaches zero (Smola and Olkopf, 2004). Figure 3.6 shows a schematic of support vector regression model.

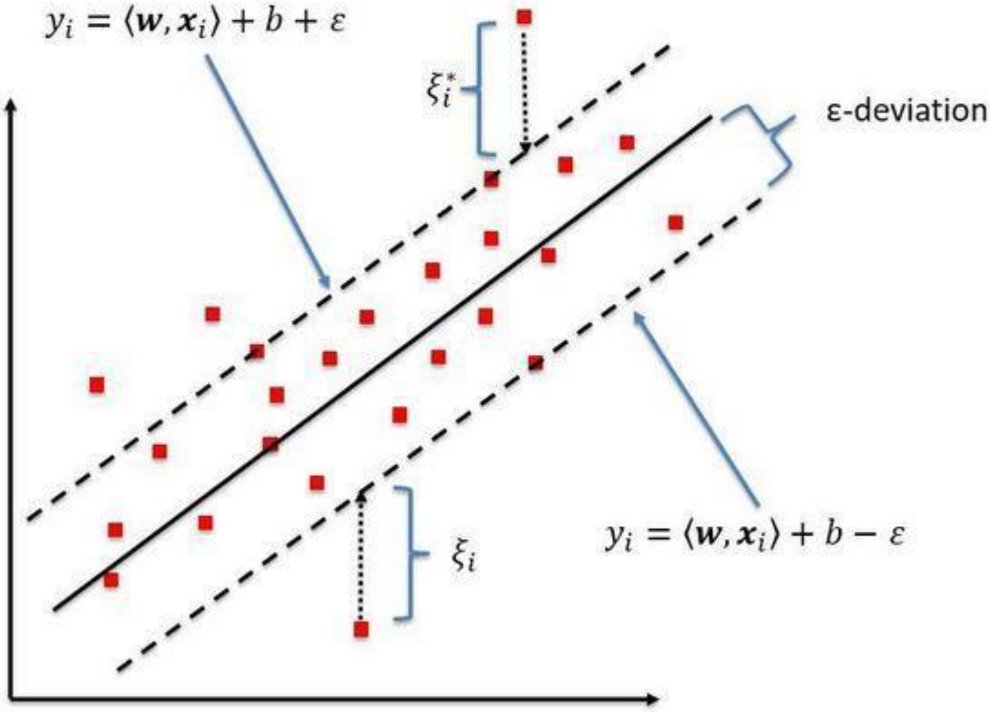

$y_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b + \varepsilon$
ξ_i^*
ε-deviation
ξ_i
$y_i = \langle \mathbf{w}, \mathbf{x}_i \rangle + b - \varepsilon$

Figure 3.6: Schematic showing a one-dimensional Support vector regression model showing hyperplane, ε – deviation and ξ is the deviation from the margin (Kleynhans et al., 2017).

Support vector regression gained popularity because it can effectively classify non-linear data by mapping data into high-dimensional feature spaces even when the datasets are small. Kernel function enables one to locate a hyperplane in the higher dimensional space without elevating computational cost. Increase in the dimension of data leads to a rise in the computational cost. When dimension increases and the separating hyperplane is not found in a particular dimension, a kernel is expected to shift the data to a higher dimension support vector classifier. It is accomplished by adding a kernel trick which maps classes into a higher dimensional space, where they are linearly separated. Kernels are classified into linear, polynomial, radial basis function kernels. Function (ϕ) maps training vectors (x_i) into higher dimensional space, this is known as the kernel trick $K(x_i, x_j)$ expressed by the equation

$$K(x_i, x_j) \equiv \phi(x_i)^T (x_j). \qquad (3.29)$$

Furthermore, SVR is less prone to overfitting issue, after training the prediction phase is rapid and work well with high dimensional data these, features make SVR more practical.

3.5.2: Random Forest Regression

Random forest regression (RFR) involves growing of trees depending on random vector Θ_k such that the tree predictor $h(x, \Theta)$ takes a numerical value. A random forest is built by taking an average over k of the trees to reduce the variance hence finding a balance between the two extremes which is expressed as (Breiman, 2001)

$$RFR = \{h(x, \Theta_k), \qquad k = 1, \ldots\} \tag{3.30}$$

where $\{\Theta_k\}$ is the random vector and $h(x, \Theta)$ is the tree predictor

Random vectors $\{\Theta_k\}$ are independently identically distributed and each tree selects the most popular class at input *x* vectors (Breiman, 2001). The mean squared generalization error (GE) for predictor is $h(x)$ is given by the equation

$$GE = E_{xy}(Y - h(x))^2 \tag{3.31}$$

where E_{xy} is expected value

The GE for forests converges as to a limit as the number of trees increases. For an accurate RFR model low correlation between residuals and low error trees are key (Breiman, 2001). The more the number of trees the more robust the forest becomes. The RFR do not over fit data as more tress are added but GE is produced.

3.6: Accuracy Metrics for the evaluation of prediction models

Several metrics are used to determine the accuracy solar (PV) prediction models based on ML techniques. They include mean squared error (MSE), coefficient of determination (R^2), Adjusted R^2 and mean absolute error. The MSE measures an average value of the squares of errors, expressed in the equation (Kim *et al.*, 2019)

$$MSE = \frac{1}{B}\sum_{i=1}^{B}\left(y_i - y_p\right)^2 = RMSE^2 \tag{3.32}$$

where y_i is the i-th actual value, $y_p$is the predicted value for y_i, B is the number of samples, and RMSE is the square root of MSE. When the RMSE decreases the predictive model's performance increases. Mean absolute error (MAE) is the average difference between the predicted and real values, it is computed using the equation

$$MAE \; = \frac{1}{B}\sum_{B}|y_i - y_p| \tag{3.33}$$

The MAE shows measure of errors between the predicted values and the real values but does not indicate the direction of the error.

Coefficient of determination (R^2) is the proportion of the variance of the dependent variable which the independent variables describe as expressed in the equation (Kim *et al.*, 2019)

$$R^2 \; = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_p\right)^2}{\sum_{i=1}^{n}(y_i - y_m)^2} \tag{3.34}$$

where y_m is the mean of the actual values of y_i.

The adjusted R-squared is a modification of R-squared and only increases when the independent variable is significant. It is expressed in the equation (Kim *et al.*, 2019)

$$Adj\ R^2 \; = 1 - (1 - R^2)\frac{B - 1}{B - p - 1} \tag{3.35}$$

where p is the total number of independent variables.

CHAPTER FOUR: MATERIALS AND METHODS

4.0: Chapter Overview

This chapter elaborates on the experimental setup and devices used. Modelling of I-V curve based on single diode equation is also explained. It further describes data preprocessing and analysis techniques used and machine learning based PV power prediction models built.

4.1: Experimental setup

The experimental setup consisted of six 250W Polycrystalline Solinc solar panels connected in series shown on Figure 4.1, weather station, HT304N standard cell, PT300N temperature sensor. Solar I-V analyzer connection cables and toolbox.

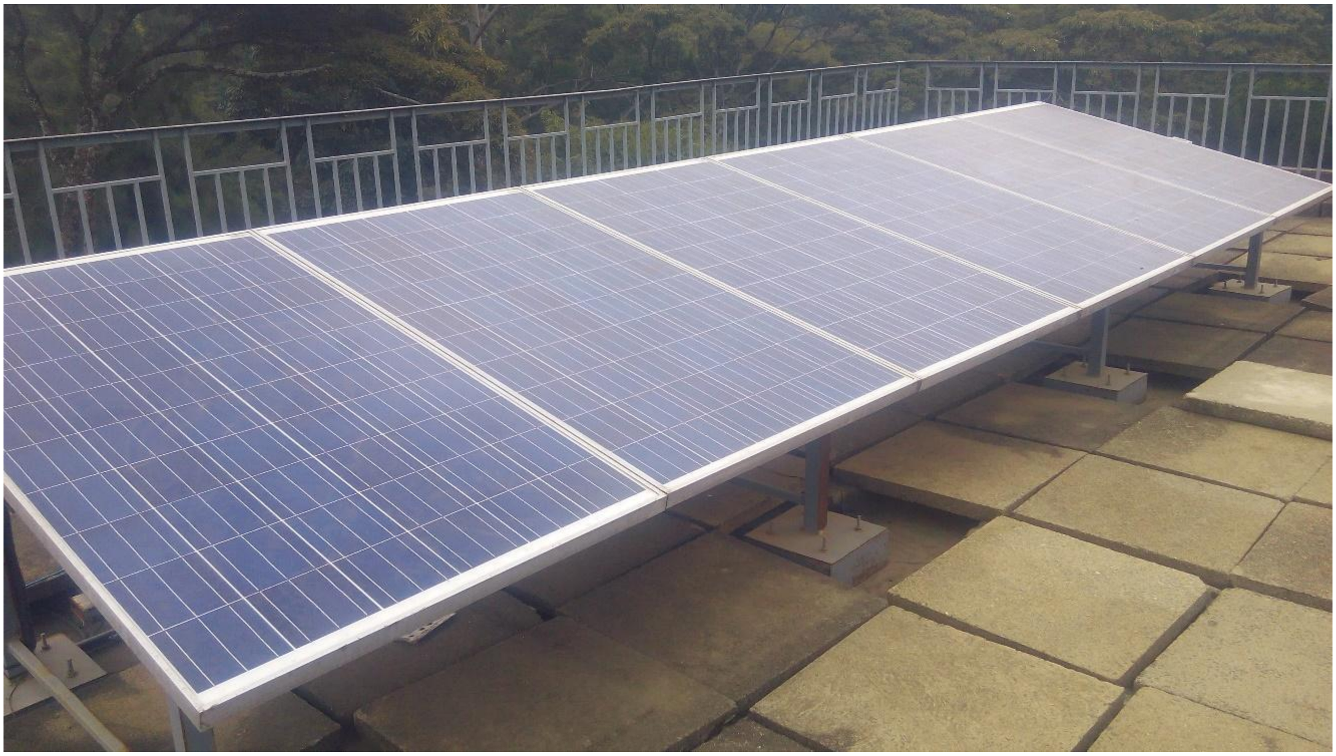


Figure 4.1: Image of the 1.5kW PV Module station

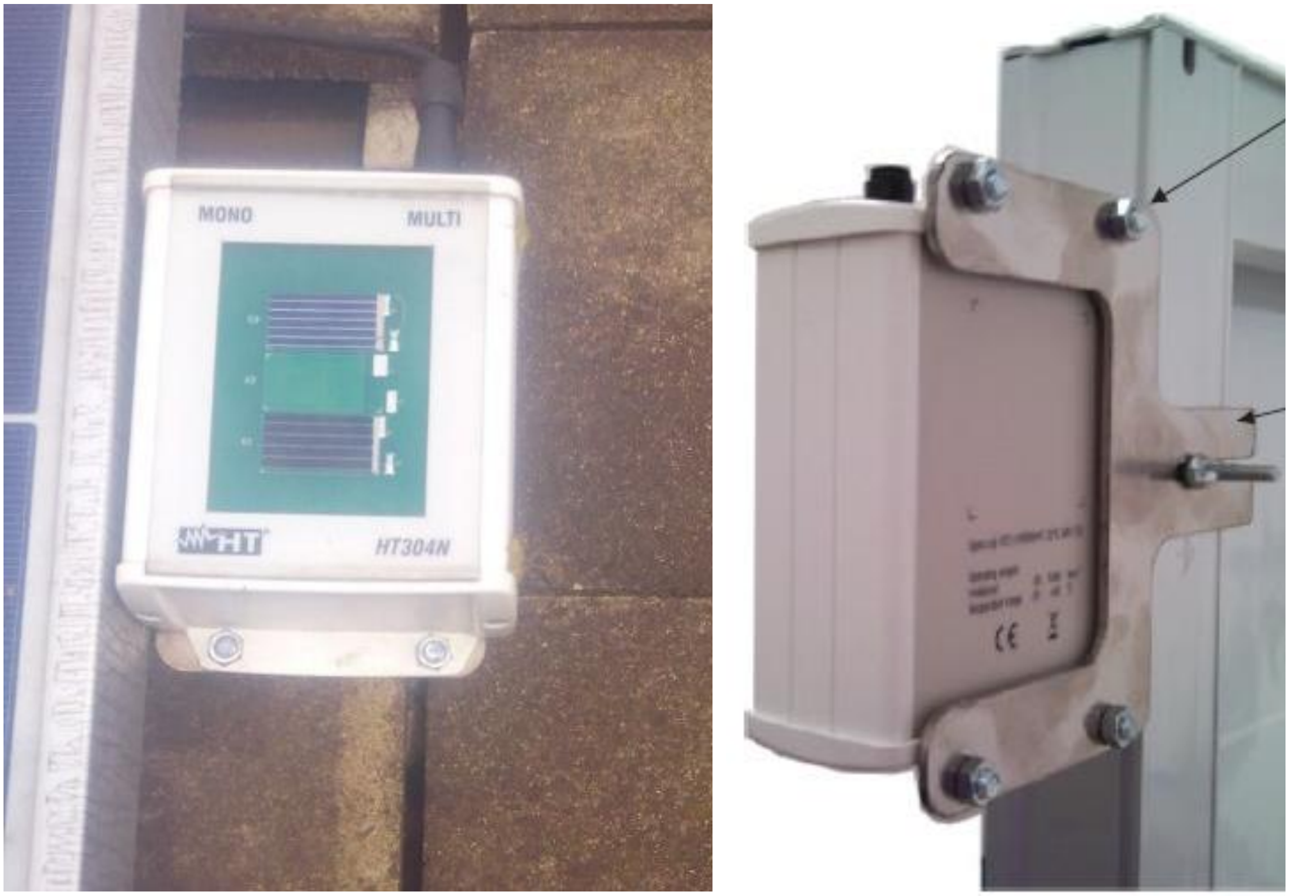
4.1.1: PV Module Station

The PV module station consisted of six 250W solar panels, HT 304N cell, PT300N temperature probe and Solar I-V analyzer. The 1.5kW PV string is installed at the School of Physical Sciences, Chiromo University of Nairobi. The solar panels were only cleaned once using a clean cloth and plain water before commencing with the measurements. Table 4.1 shows the technical specifications of the PV module at STC as provided by the manufacturer.

Table 4.1: The technical specifications of the 250W solar module used in the experiment at Standard testing conditions (Solinic East Africa.Limited, 2017).

Maximum Power	250W
Power Tolerance	+/- 3%
Current at maximum power point	8.6A
Voltage at Maximum power point	29V
Open Circuit voltage	36V
Short Circuit Current	9.5A
Nominal Operating Temperature	46 +/-°C
Operating Temperature	-40°C-80°C
Temperature Coefficient for Maximum Power	-0.42%/°C
Temperature Coefficient for open circuit voltage	-0.32%/°C
Temperature Coefficient for short circuit current	0.05%/°C
Cell Efficiency	>17.6%
Solar module Area	1.4415m^2
Solar Cell Area	240.25cm^2

The solar irradiance was measured using a HT304N reference cell which was first cleaned using water and a clean cloth to ensure the accuracy of its readings. The cell has a capacity to measure irradiance ranging from 50-1400W/m^2 with an accuracy of ±3% (Italia, 2012). It has two output ports one for Si-Multi Crystalline cell and another for Si-Mono Crystalline cell, the multi crystalline output port was used for this study as shown in Figure 4.2 (a). The HT304N cell was attached to the solar panel frame as shown in Figure 4.2 (b) using a stirrup with a fixing screw for ease of installation and to ensure that it has the same orientation and inclination as the solar panels.



(a) **(b)**

Figure 4.2: Front view of the HT304 Standard Cell with two Si-Cells (a), Back view of the HT304N attached to the back frame of PV module using the stirrup and screws (b).

The module temperature was measured using the temperature Sensor (PT300N) with a capacity to measure temperature from -20°C to 99.9°C with an uncertainty value of (± 1%). The PT300N was attached to the back of one solar module using adhesive tape as shown in the Figure 4.3 and its output probe connected to the Solar I-V Analyzer.

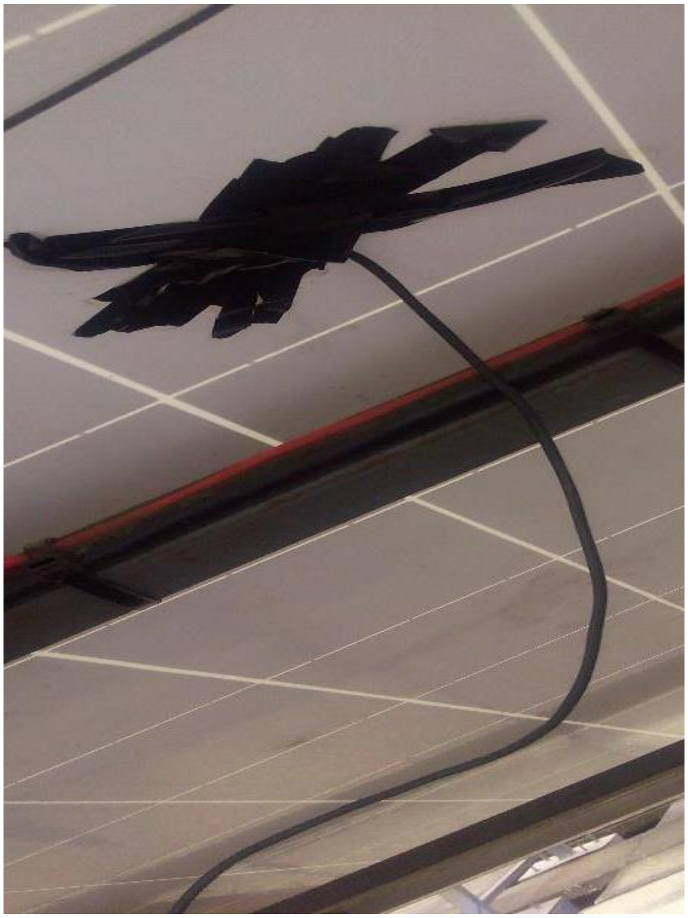


Figure 4.3: Temperature sensor PT300N attached with adhesive tape at the back of a solar module.

Solar current-voltage analyzer was used to measure the power output of the PV string this was done by connecting positive and negative connection cables of the PV string to the Solar I-V Analyzer. Figure 4.4 shows an image of a Solar I-V Analyzer.

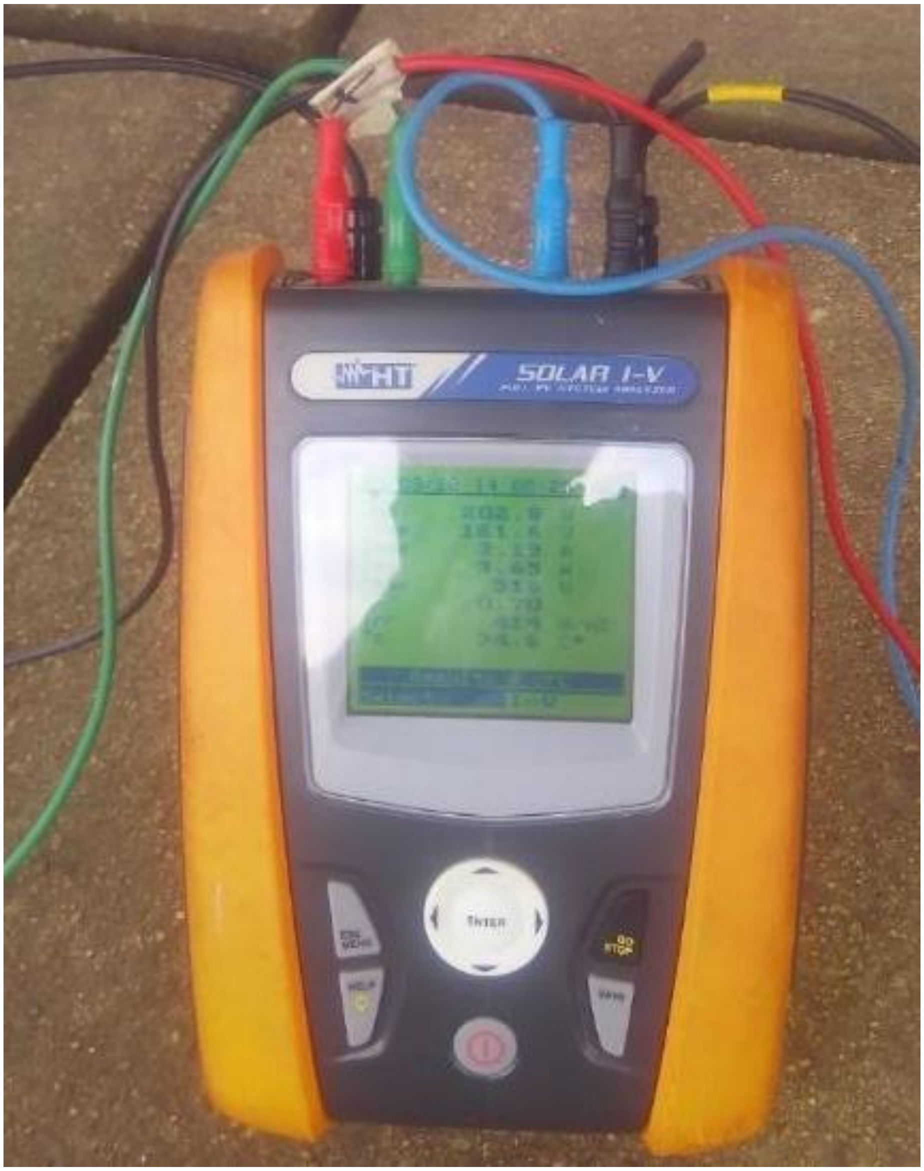
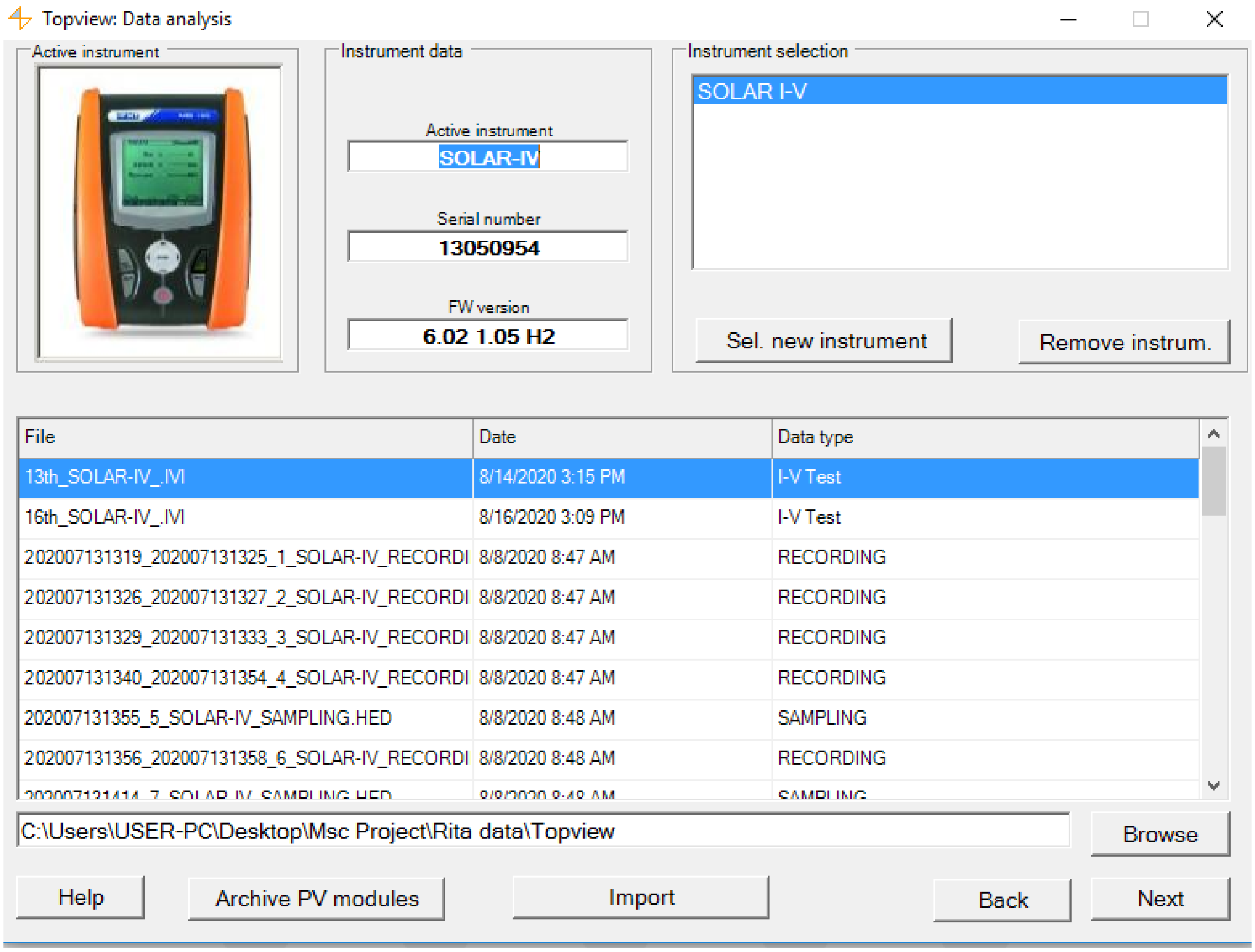


Figure 4.4: Solar Current –Voltage (I-V) 400 analyzer from HT instruments.

Solar I-V analyzer was also used to log the data from the HT304N cell and PT300N probe respectively. The collected data obtained was extracted from the Solar I-V analyzer to a computer using C2006 optical cable and finally displayed on the Top View software as displayed by Figure 4.5 .



Topview: Data analysis

Active instrument

Instrument data

Active instrument: SOLAR-IV

Serial number: 13050954

FW version: 6.02 1.05 H2

Instrument selection

SOLAR I-V

File	Date	Data type
13th_SOLAR-IV_.IVI	8/14/2020 3:15 PM	I-V Test
16th_SOLAR-IV_.IVI	8/16/2020 3:09 PM	I-V Test
202007131319_202007131325_1_SOLAR-IV_RECORDI	8/8/2020 8:47 AM	RECORDING
202007131326_202007131327_2_SOLAR-IV_RECORDI	8/8/2020 8:47 AM	RECORDING
202007131329_202007131333_3_SOLAR-IV_RECORDI	8/8/2020 8:47 AM	RECORDING
202007131340_202007131354_4_SOLAR-IV_RECORDI	8/8/2020 8:47 AM	RECORDING
202007131355_5_SOLAR-IV_SAMPLING.HED	8/8/2020 8:48 AM	SAMPLING
202007131356_202007131358_6_SOLAR-IV_RECORDI	8/8/2020 8:48 AM	RECORDING
202007131414_7_SOLAR-IV_SAMPLING.HED	8/8/2020 8:48 AM	SAMPLING

C:\Users\USER-PC\Desktop\Msc Project\Rita data\Topview

Figure 4.5: An image of the Top View Software displayed on Computer monitor when connected to the Solar I-V 400 analyzer.

4.1.2: Weather station setup

A portable weather station equipped with an anemometer, thermo-hygrosensor, wind direction and rain gauge was utilized to measure relative humidity and ambient temperature. It was calibrated as per the local time settings for the ease in the radio-controlled transmission of the data to a work station computer. The data obtained was logged real time from 10:00 AM to 3:00 PM at 30 minute's intervals for a period of three weeks. The weather station used for this study is shown in the Figure 4.6

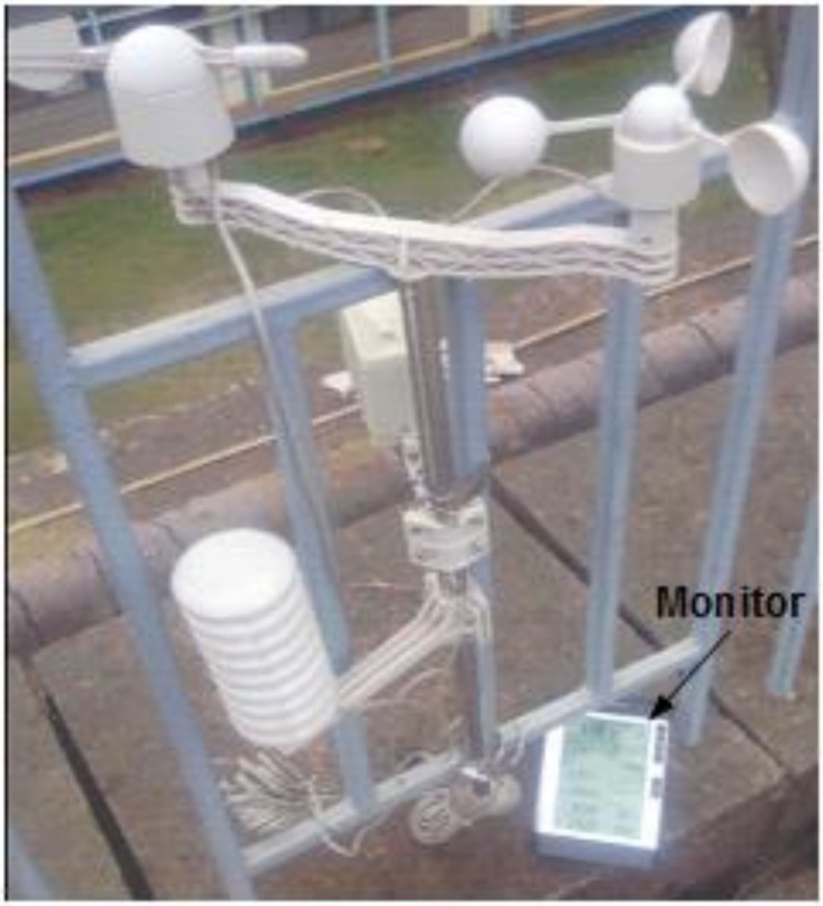


Figure 4.6: A picture of the portable Weather station installed at the School of Physical sciences, University of Nairobi the radio-controlled monitor displaying real time weather data is also shown.

4.1.3: Curve modelling

I-V curve model based on the single diode equation was built using shiny application in R-software environment. The shiny application consisted of a user interface and a server. The user interface (UI) was used to create an interactive front end that prompts a user to input short circuit current, open circuit voltage, module temperature and irradiance. Data obtained from the user is fed into the server where it is the single diode model equations are computed resulting into I-V and P-V curve. Shiny app code is documented in Appendix 4.

4.2: Data preprocessing and analysis

In order to do data preprocessing, exploratory data analysis was first conducted on the raw data collected in this work. The structure of the data was first established by determining the variable type, number of variables and observations by loading the data in R script. The variable type was also determined. Checking for missing data was done to ensure the completeness of the data. The minimum, maximum mean, median, lower and upper quartile of the dataset was calculated using the in-built R command summary, the distribution of the raw data was plotted in Origin 9.0 software to discover any trends and outliers in the data collected. The scatter plots were used as a basis for identifying relationships between various variables with the correlation between ambient temperature, module temperature, relative humidity and irradiance done using R script in Appendix 1 based on the Pearson's Correlation Coefficient expressed in equation (3.16). Prior to performing PCA the data was centered and scaled, PCA was then done on the data to reduce dimensionality using the prcomp command in R Software as documented in Appendix 1

4.3: Machine learning (ML)

The data obtained from the PCA was divided into training, testing and validation dataset using the ratio 60%, 20% and 20% respectively. Cross validation (CV) was done using the *k-fold* CV, *leave one out* CV and random resampling CV on the training dataset, in order to prevent the models from overfitting see in Appendix 2. The training dataset was used to build the Support Vector regression (SVR) with radial basis kernel model in the R software. The penalty parameter C was set at 0.25, 0.5 and 1, the sigma was set at 0.05. The model was built and accuracy metrics computed as documented in Appendix 2. The validation data was then utilized to evaluate the performance of the model. The training dataset was used to build the Random Forest Regression (RFR) model using R Software. 20 trees were used in the model. The model's accuracy metrics for the model were also computed as documented in Appendix 3. An interactive shiny application based on the best predictive model was designed to enable prediction of the PV power output when inputs irradiance, relative humidity, ambient temperature and module temperature are fed into the app. Figure 4.7 shows a diagram summarizing the key techniques utilized in this study.

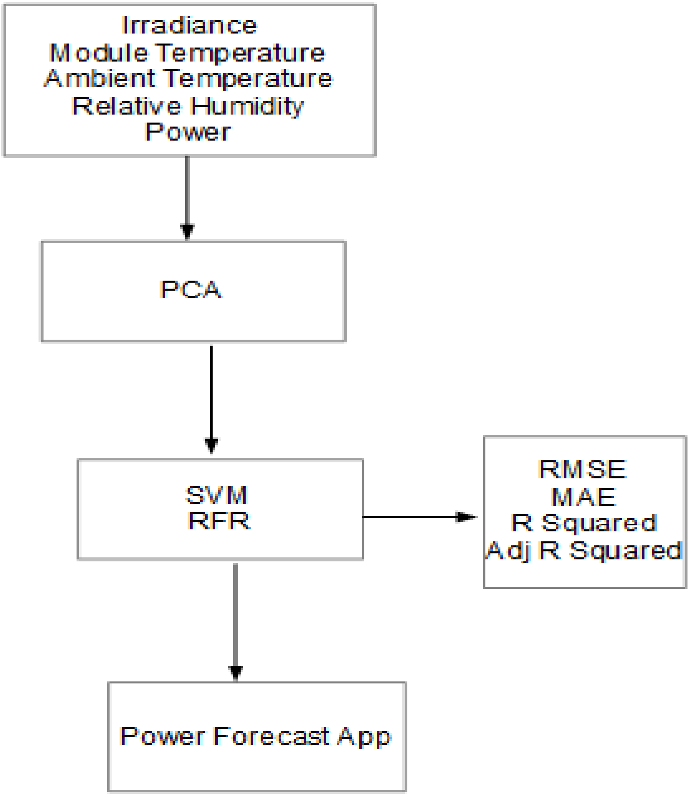


Figure 4.7: A Block diagram showing the different Machine learning techniques used in this study.

CHAPTER FIVE: RESULTS AND DISCUSSION

5.0: Chapter Overview

In this chapter, the effect of solar irradiance, relative humidity and module temperature on the performance of the 1.5kW PV system is discussed. This chapter further analyses the performance of PV power output predictive models built using SVR and RFR, and a PV power forecast app is also built. An application based on the single diode model is built and its performance evaluated.

5.1: Variation of weather parameters within three weeks

Exploratory data analysis was done to discover trends and relationships between 231 observables and 5 variables collected for a period of three weeks shown in Appendix 6. Table 5.1 shows the summary of the maximum power, solar irradiance, relative humidity, module and ambient temperature. The summary incorporates the minimum, maximum and mean value of the variables. The 1st Quartile represents the midway value obtained between the minimum value and the median of the variables. The median (2nd Quartile) is the midway of the data while 3rd quartile is the midway value between the median and maximum value of the dataset. The Figure 5.1 demonstrates this clearly.

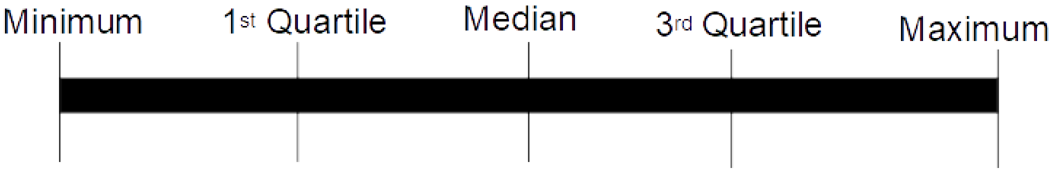


Figure 5.1: Illustration of minimum, 1st quartile, median, 3rd quartile and maximum .

Table 5.1: Summary of power output, solar irradiance, module temperature, ambient temperature and relative humidity for the three weeks.

Summary of the data collected	**Maximum Power (W)** **+ 0.1**	**Solar Irradiance (W/m^2)** **+ 0.1**	**Module Temperature (°C)** **+ 0.1**	**Ambient Temperature(°C)** **+ 0.1**	**Relative Humidity (%)** **+ 0.1**
Minimum	73.0	51.9	15.5	14.6	29.0
1st Quartile	319.5	246.5	27.8	20.6	45.0
Median (2nd Quartile)	478.0	385.0	33.6	22.7	54.0
Mean	542.9	468.3	33.1	22.7	56.3
3rd Quartile	791.5	692.0	39.3	25.4	64.5
Maximum	1209.0	1117.5	50.3	30.4	98.0

The variation of the average daily solar irradiance, relative humidity and ambient temperature for three weeks was studied. Figure 5.2, Figure 5.3 and Figure 5.4 shows the average daily solar irradiance, relative humidity and ambient temperature respectively. In week one (day 1-7) there was relatively low relative humidity, high solar irradiation, and high ambient temperature. In week two (day 8-14), the solar irradiance and ambient temperature decreased significantly, while the relative humidity began to increase gradually as the weather shifted to slightly cloudy, chilly, and windy conditions. The third week (day 14-21) a large drop in solar irradiance and ambient temperature was observed, though humidity remained high; these conditions were due to the cloudy, light rain, and windy weather experienced.

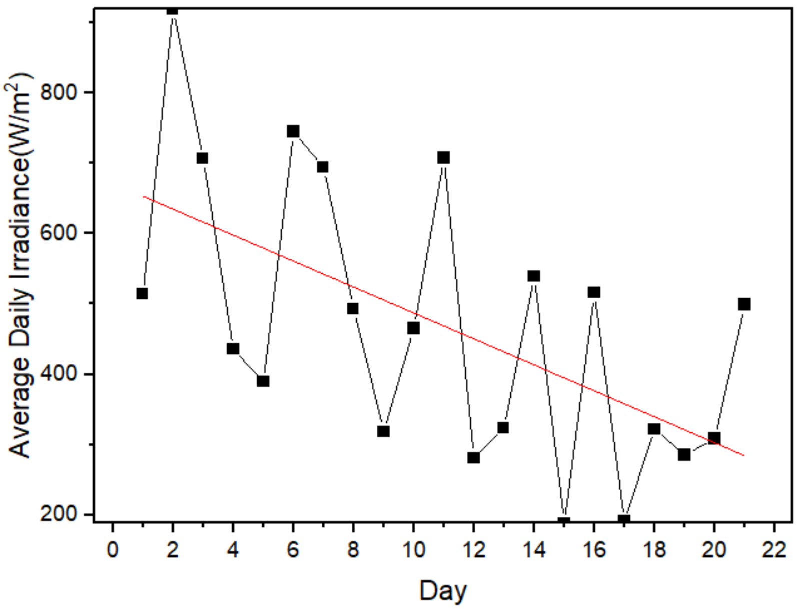


Figure 5.2: Variation of average daily solar irradiance with the day

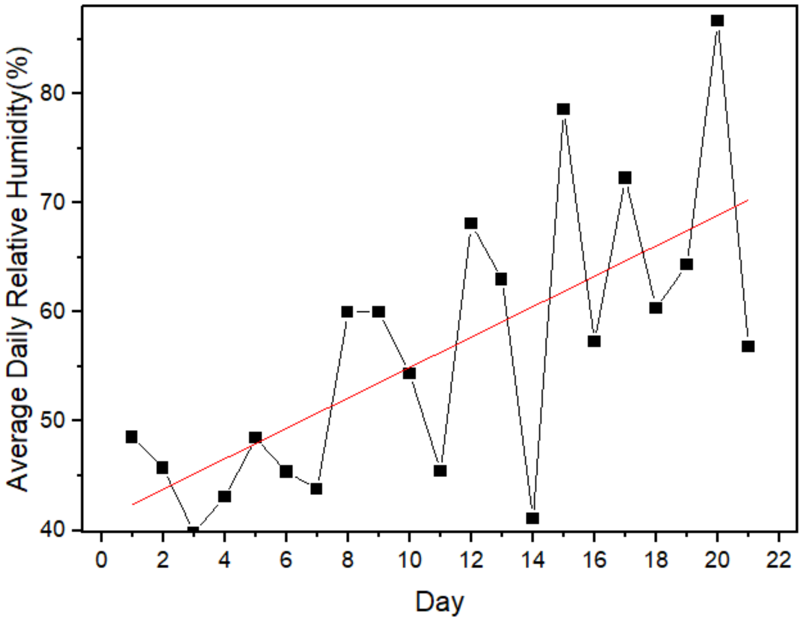


Figure 5.3: Variation of average daily relative humidity with the day.

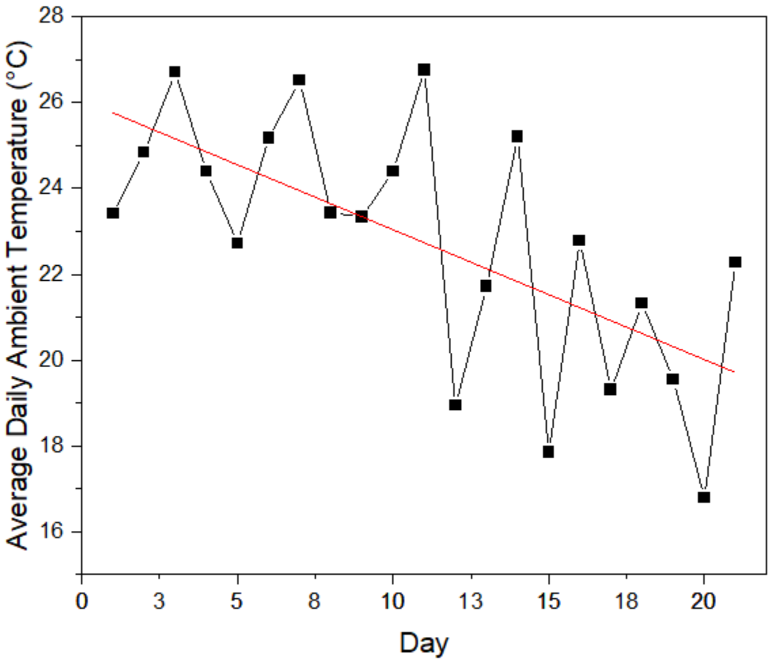


Figure 5.4: Variation of average daily ambient temperature with the day.

5.1.1: The variation of ambient temperature with solar irradiance

Figure 5.5 shows the variation of ambient temperature (T_a) and solar irradiance (G_a). It is noted that the ambient temperature increases weakly with solar irradiance. Maximum ambient temperature recorded was 30.4°C at a solar irradiance of 906 W/m^2 while the lowest ambient temperature was recorded at a 14.6°C at solar irradiance of 52 W/m^2. Data points are scattered from the line of fit and spread all over the plot showing weak relationship between the variables with the Adj R^2 of 0.27. This is confirmed by a positive coefficient (P) of 0.53 between the variables. This relationship can be credited to the fact that the earth's surface readily absorbs solar irradiance. and converts it to heat energy. The earth re-emits the heat which is then trapped within the earth's atmosphere resulting in a rise in the ambient temperature. A similar pattern was observed by Shrestha *et al*. (2019) who reported a correlation coefficient (P) of 0.9055 between ambient temperature and solar irradiance for data collected for one year which is much higher than p of 0.53 acquired for data of three weeks.

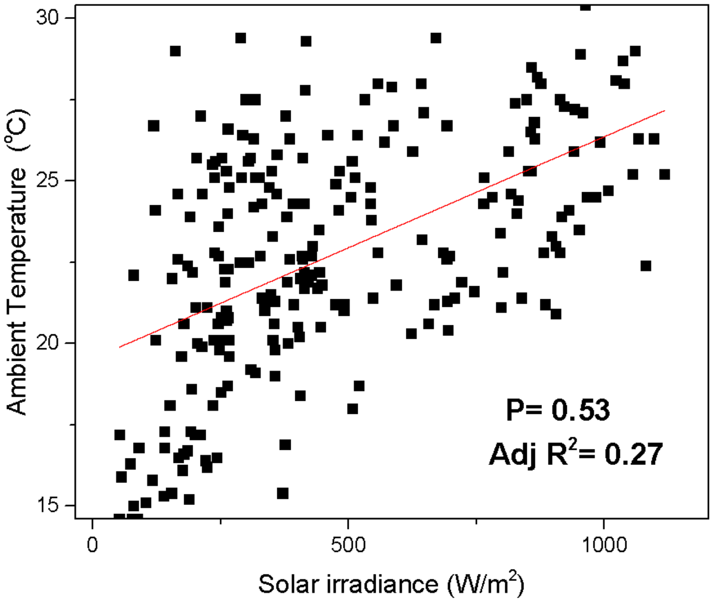


Figure 5.5: Variation of ambient temperature with solar irradiance.

This linear relationship between ambient temperature (T_a) and solar irradiance (G_a) and can be expressed in equation 5.1.

$$T_a = 0.00683(G_a) + 19.54 \qquad (5.1)$$

5.1.2: The variation of solar irradiance with relative humidity

Figure 5.6 shows the relationship of solar irradiance and relative humidity. A maximum relative humidity of 98% was recorded at a solar irradiance of 88W/m^2 while the lowest relative humidity was obtained 29% at a solar irradiance of 289 W/ m^2. An increase in the amount of solar irradiance results in a decrease in the relative humidity depicting inverse relationship confirmed by negative correlation coefficient (P) of- 0.51. However, the relationship is weak especially at relative humidity lower than 65% with Adj R^2 of 0.25. This trend can be attributed to the decrease in the amount of water vapor in the atmosphere experienced when the ambient temperature increases due to the increase in the amount of solar irradiance.

Nicholas *et al.*, (2018) investigated the relationship of relative humidity and solar irradiance and reported an inverse relationship.

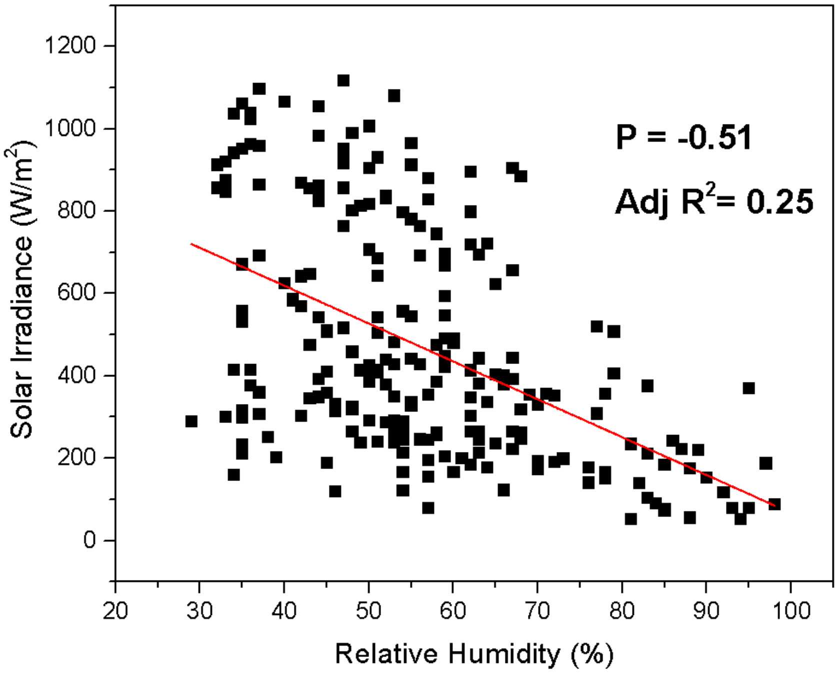


Figure 5.6: Variation of solar irradiance with relative humidity.

The relationship between solar irradiance and relative humidity is best expressed by linear equation 5.2

$$Hum = -9.1983(G_a) + 986.50 \tag{5.2}$$

5.1.3: The variation of ambient temperature with relative humidity

Figure 5.7 shows the relationship between ambient temperature and relative humidity. The maximum ambient temperature of 30.4°C was recorded at relative humidity of 36% while lowest ambient temperature was 14.6°C at relative humidity of 85%. It was noted that ambient temperature increased with decreasing relative humidity indicating an inverse relationship confirmed by the high negative correlation coefficient (P) of -0.95 computed. The data points are closely clustered along the red line of fit indicating a strong relationship with Adj R^2 of 0.90. This relationship can be attributed to the increase in the ambient temperature that leads to a decrease in the amount of water vapor in the atmosphere hence decreases the relative humidity. A similar trend was observed by Wu and Lou, (2011).

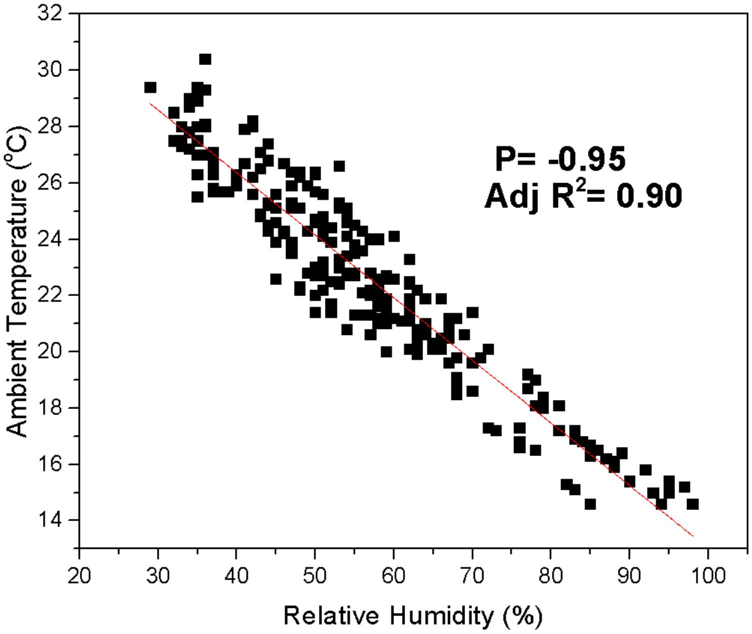


Figure 5.7: Variation of ambient temperature with relative humidity.

The relationship between ambient temperature (T_a) and relative humidity(Hum) can be expressed by equation 5.3

$$T_a = -0.22244(Hum) + 35.26 \tag{5.3}$$

5.1.4: The variation of ambient temperature with module temperature

Figure 5.8 shows the relationship between ambient temperature and module temperature. The maximum module temperature was recorded at 50.3°C while the ambient temperature was at its highest of 30.4°C. The module temperature was found to be higher than the ambient temperature, with a 10.4°C difference on average. It was established that the module temperature increases with increasing ambient temperature indicating a positive correlation with a correlation coefficient (P) of 0.85. From Figure 5.8 it can also be noted that the data points spread further from the regression line at ambient temperatures higher than 21°C indicating a relatively good fit of Adj R^2 of 0.72. The module temperature is higher than the ambient temperature because both the thermal energy emitted by the solar panel during the photovoltaic phase and the absorbed heat by the PV module from the irradiance falling contribute to the module's temperature. Ciulla *et al.*, (2013) and Bashir *et al.*, (2014) work discovered a similar trend.

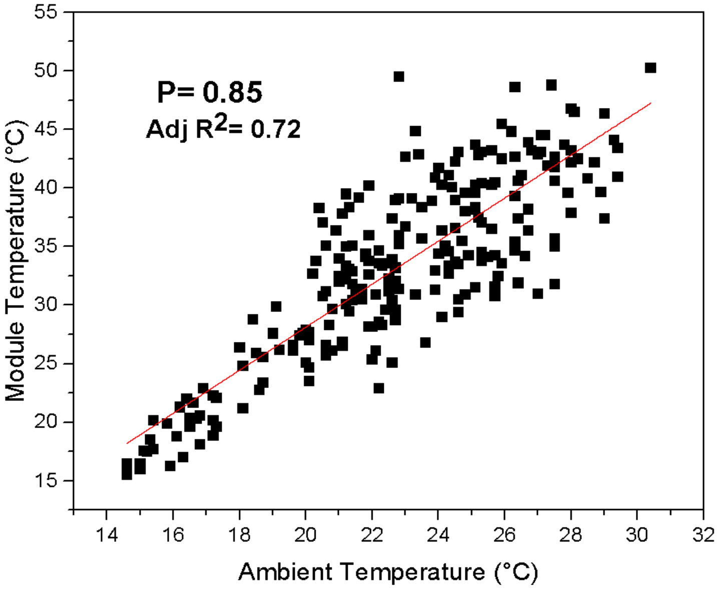


Figure 5.8: Variation of module temperature with ambient temperature.

The linear fit between ambient temperature and module temperature is best expressed by equation 5.4

$$T_m = 1.84(T_a) - 8.65 \tag{5.4}$$

5.2: Effect of solar irradiance on the performance of the PV system

This section outlines the effect of solar irradiance on the performance of the 1.5kW PV system. It highlights the effect of solar irradiance on the maximum power output, short circuit current, series resistance, shunt resistance and the PV system efficiency.

5.2.1: Effect of solar irradiance on the maximum power

Figure 5.9 shows the relationship between maximum PV power output and solar irradiance. The maximum power output was highest at 1209W at a solar irradiance of 1081W/m ^{2} while the lowest power was measured at 73W at a solar irradiance of 52W/m^2. It is noted that the maximum power linearly increases with solar irradiance indicating high correlation confirmed by (P=0.99). The data points are clustered along regression line and strong linear relationship demonstrated by (Adj R^2 of 0.98) indicating a good linear fit. This relationship can be credited to the increase in the rate of photo generation as solar irradiance increased consequently leading to an increase in the power output, similar results were obtained by Omubo-Pepple *et al.*, (2009) and Musanga *et al.*, (2018).

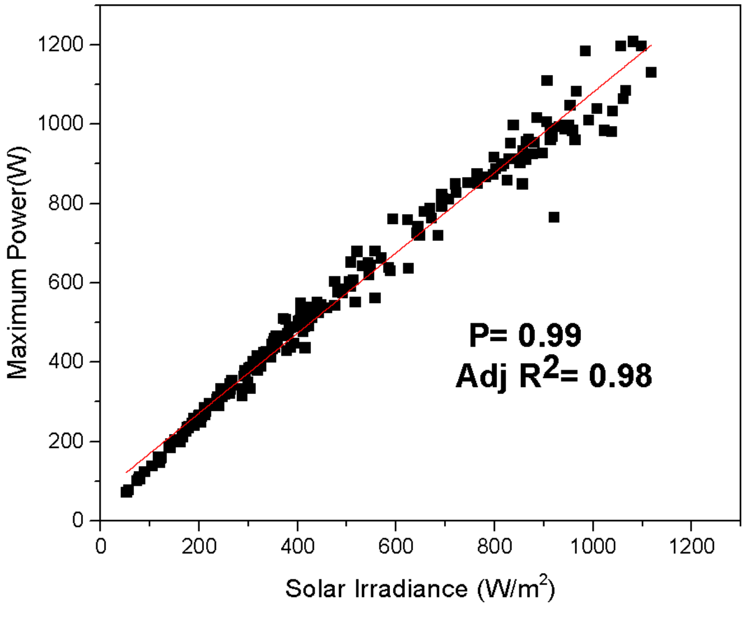


Figure 5.9: Variation of the maximum PV power with solar irradiance.

The relationship between the maximum PV power (P_m) and solar irradiance (G_a) is expressed using equation which is obtained from a linear model expressed in equation 5.5

$$P_m = 1.01(G_a) + 69.45 \tag{5.5}$$

5.2.2: Effect of solar irradiance on the short circuit current

Figure 5.10 shows the variation of the short circuit current with solar irradiance. The highest short circuit current recorded was 9.88A at a solar irradiance of 1117.5W/m^2 while lowest was at 0.5A at a solar irradiance of 52W/m^2. The short circuit current was observed to increase with increasing solar irradiance which was confirmed high positive correlation (P = 0.99) and a good linear fit of Adj R^2 =0.98. This strong relation can be attributed to the increase of photo electrons generated as a result of the increase in the amount of solar irradiance falling on the PV modules consequently increasing the amount of current generated by PV modules. A similar observation was obtained by Omubo-Pepple *et al.*, (2009), and Musanga *et al.*, (2018).

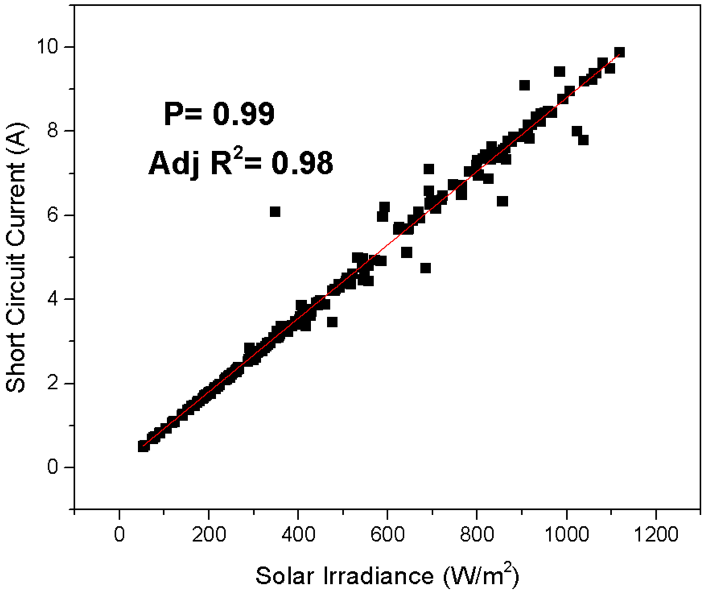


Figure 5.10: Variation of short circuit current with solar irradiance.

The relationship between the short circuit current (I_s) and irradiance (G_a) is expressed using equation 5.6 which is obtained from a linear fit

$$I_s = 0.009(G_a) + 0.06 \tag{5.6}$$

5.2.3: Effect of solar irradiance on the series resistance

Figure 5.11 shows the relationship between series resistance with solar irradiance. The series resistance is highest at 37 Ω at an irradiance of 51.9 W/m^2 while the lowest series resistance was recorded at 4.46 Ω at a solar irradiance of 1096.8 W/m^2. The figure shows an exponential decay curve. The data points lie on the logistic curve showing perfect fit of Adj R^2 =1. It is observed that there a rapid decrease in series resistance a low irradiance (<100W/m^2) and thereafter starts to flatten. This behavior can be attributed to the increase in the electron-hole generation as the solar irradiance increased which leads to increase in the conductivity consequently decreases the series resistance, Benda and Machacek, (2016) observed a similar relationship.

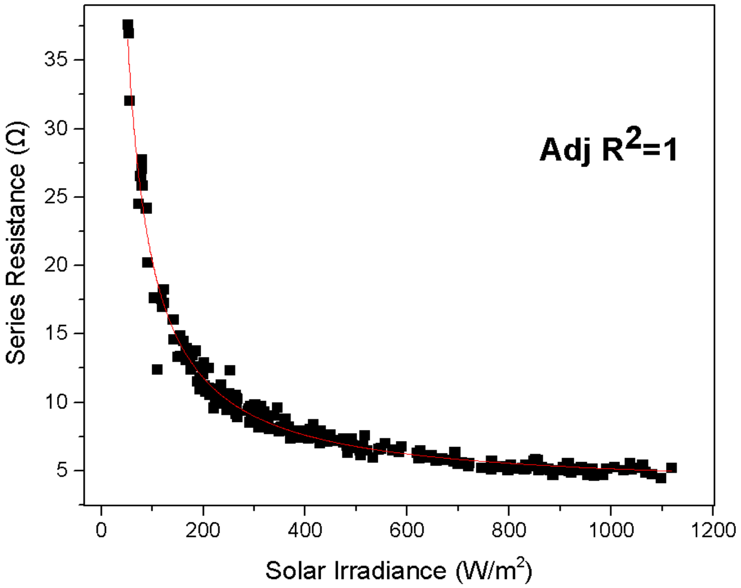


Figure 5.11: Variation of series resistance with solar irradiance.

This trend can be described by the logistic fit equation 5.8 of the form

$$R_s = 3.59 + \frac{(33401.6 - 3.59)}{1 + \left(\frac{G_a}{0.063}\right)^{1.03}} \tag{5.8}$$

5.2.4: Effect of solar irradiance on the shunt resistance

Figure 5.12 shows the behavior of solar irradiance with shunt resistance. The figure shows an exponential decay curve where the shunt resistance decreased with increase in solar irradiance. The data points lie along curve but outliers away from the fit indicate a negative weak non-linear relationship of Adj R^2 computed to 0.64. This relationship is due to increases in free charge carriers when solar irradiance increases leading to high conductivity hence leads to low shunt resistance. Similar trend was discovered by Benda and Machacek, (2016).

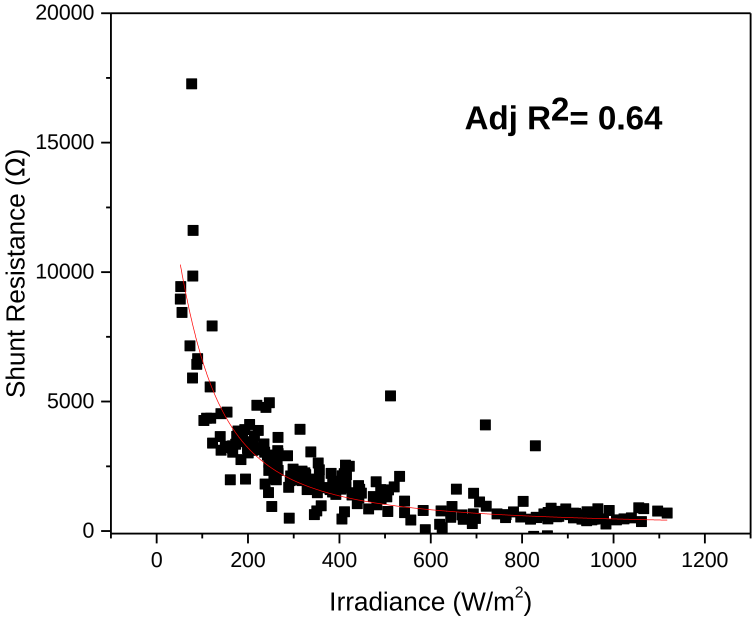


Figure 5.12: Variation of shunt resistance with solar irradiance.

The relationship can be best described by the logistic fit equation 5.9 of the form

$$R_p = 162.98 + \frac{(15096.7 - 162.98)}{1 + \left(\frac{G_a}{83.66}\right)^{1.56165}} \tag{5.9}$$

5.2.5: Effect of solar irradiance on the PV system efficiency

The relationship of module efficiency with solar irradiance shown by figure 5.13, with the highest irradiance was obtained at 1117.5 W/m^2 corresponding to PV system efficiency 11.7%, while the lowest irradiance of 51.9W/m^2 corresponded to the PV system efficiency 16.18%. The PV system efficiency varied inversely with solar irradiance with negative correlation coefficient of -0.85. At low solar irradiance (<200W/m^2) the module efficiency is closely clustered along the line of fit but at high solar irradiances the data points are spreading away from the line of fit showing that module efficiency varied more randomly at high irradiance. Adj R^2 of 0.72 was obtained indicating a fairly good fit. This observation can be attributed to increase in the module temperature which increases the rate of recombination as more irradiance falls on the PV module thus significantly reducing the power output hence decreasing the module efficiency. Bashir *et al.*, (2014) and Musanga *et al.*, (2018) also reported decreasing trend of module efficiency with solar irradiance.

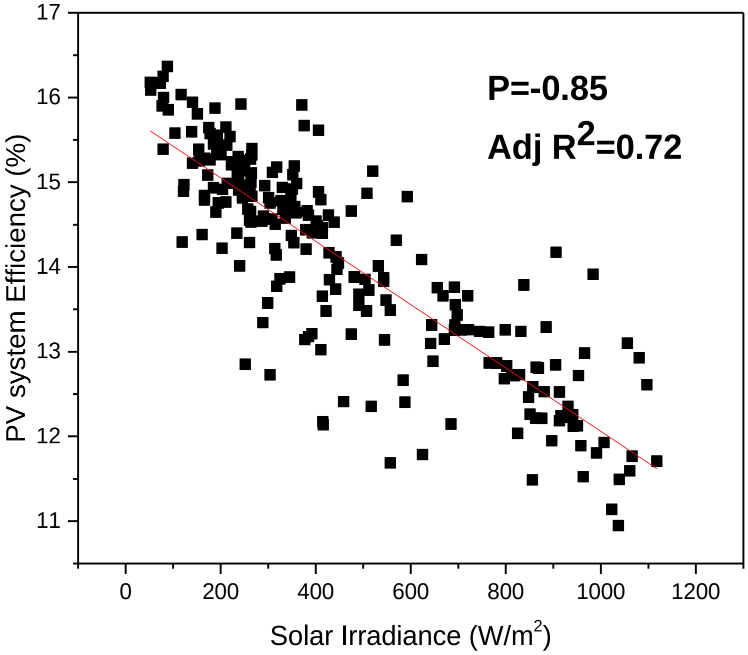


Figure 5.13: Variation of PV system efficiency with solar irradiance.

The relationship between module efficiency(η) and solar irradiance $(G_a)$is best expressed by equation 5.10

$$\eta = -0.003(G_a) + 15.79 \tag{5.10}$$

5.3: Effect of module temperature on performance of the PV system.

This section outlines the effect of module temperature on the performance of the 1.5kW PV system. It highlights the effect of the module temperature on the maximum power output, open circuit voltage, and the PV system efficiency.

5.3.1: Effect of module temperature on the maximum power

The effect of temperature on PV module performance was studied. Figure 5.14 shows variation of power output (P_m) with module temperature(T_m). The highest value of maximum power recorded was 1209W at module temperature of 33.6°C. The maximum power output varied linearly with the module temperature with yielding a positive correlation coefficient of 0.71. The scatter plot shows the data points spread from the line of fit especially at module temperature higher than 25°C showing a weak relationship (Adj R^2 of 0.49). This trend was as a result of the increase of the module temperature as the solar irradiance increases hence increasing the power output. Similar observation was observed by Musanga *et al.*, (2018).

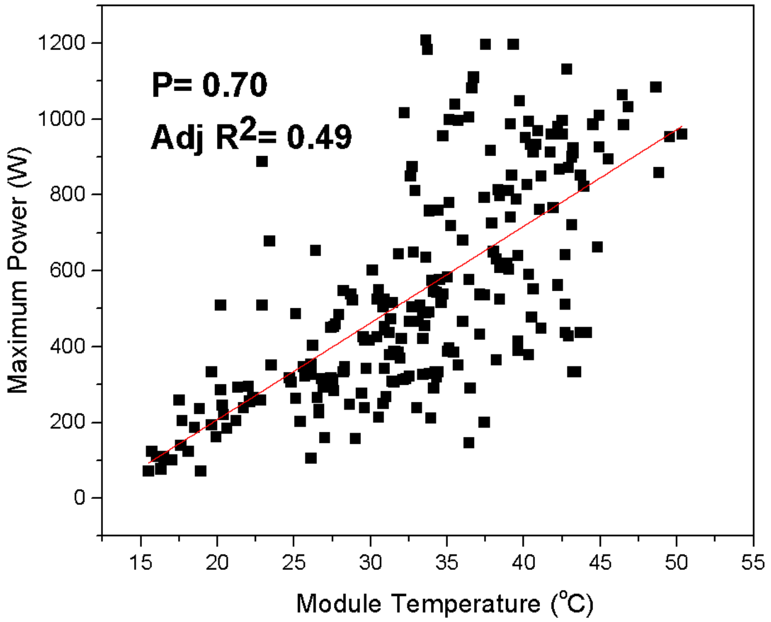


Figure 5.14: Variation of maximum power with module temperature.

The relation is expressed in the linear model expressed by equation 5.11

$$P_m = 25.51(T_m) - 302.84 \tag{5.11}$$

5.3.2: Effect of module temperature on the open circuit voltage

Figure 5.15 shows the variation of open circuit voltage (V_o) with module temperature (T_m). The maximum value of open circuit voltage obtained was 214.5V at module temperature of 20.2℃ while lowest value of open circuit voltage was 188.5V corresponding module temperature of 43.2℃. It was noted that the open circuit voltage decreases with increasing module temperature with correlation of -0.50. The data points from the figure are scattered randomly from regression line indicating a weak relationship with poor fit Adj R^2 of 0.25.

This relationship can be attributed to the increase in the rate of phonon vibrations due to thermally generated electrons produced as the module temperature increases thus disrupting the generation of electron holes, consequently leading to a decrease in the open circuit voltage. Similar results were obtained by Arjyadhara and Chitralekha, (2013), and Musanga *et al.*, (2018).

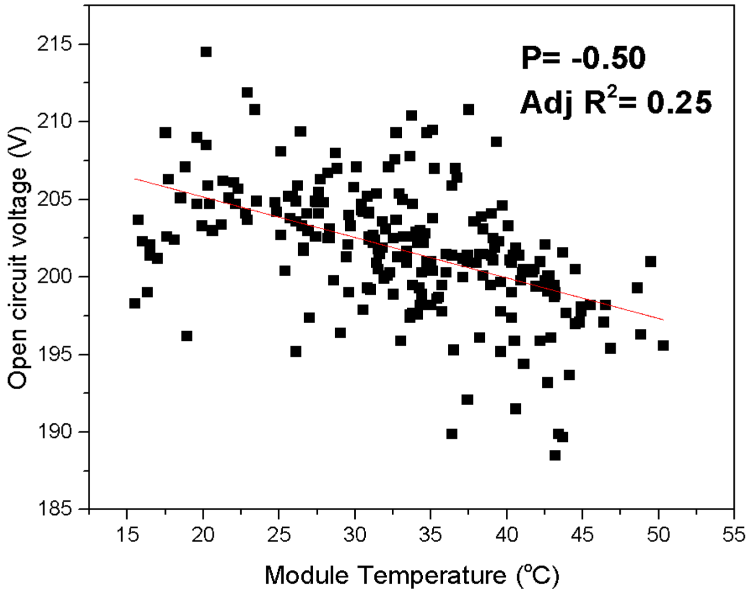


Figure 5.15: Variation of open circuit voltage with module temperature.

The relationship between open circuit voltage and module temperature is shown in figure 5.15 is best expressed by the equation 5.12 obtained from a linear fit

$$V_o = -0.26(T_m) + 210.37 \tag{5.12}$$

5.3.3: Effect of module temperature on the PV System efficiency

Figure 5.16 shows the variation of PV system efficiency (η) with module temperature. From figure it is noted that at low module temperature the PV system efficiency was high. The highest module temperature of 50.3°C the PV system efficiency computed was 11.52% while for the lowest module temperature of 15.5°C the PV system efficiency computed was 16.2%. It is noted that module efficiency decreases with increasing module temperature showing a negative linear relationship with the correlation coefficient of -0.87.

The data points clustered along the regression line showing a good fit of Adj R^2 of 0.76. The relationship results from to the increase in phonon vibrations due to thermally generated electrons produced as the module temperature increases thus disrupting the generation of electron holes, consequently decreases the open circuit voltage. Similar results were obtained by Arjyadhara and Chitralekha, (2013), Whitaker *et al.*, (1992) and Musanga *et al.*, (2018).

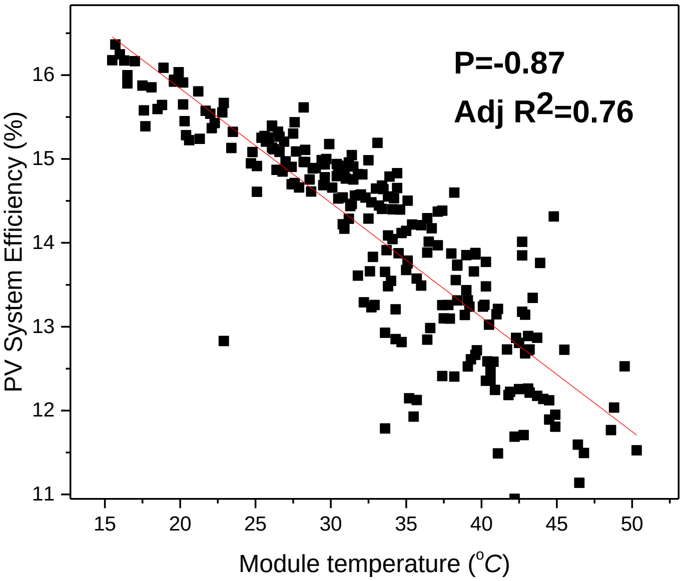


Figure 5.16: Variation of PV system efficiency with module temperature.

The relationship between PV system efficiency with module temperature is expressed in equation 5.13 of best fit

$$\eta = -0.14(T_m) + 18.57 \qquad (5.13)$$

This relationship is due to the decrease in the power as the voltage drops due to the rise in temperature ,hence reducing the PV system efficiency.

5.4: Effect of relative humidity on the performance of a PV system

This section outlines the effect of relative humidity on the performance of the 1.5kW PV system. It highlights the effect of relative humidity on the maximum PV power output and on module efficiency.

5.4.1: Effect of relative humidity on the maximum power output

Figure 5.17 shows variation of maximum power output with relative humidity. A maximum power of 1209W was recorded at relative humidity of 53% while minimum power was at 73W at relative humidity of 94%. This relationship shows that maximum power output (P_m) increases with decreasing relative humidity (Hum) indicating a negative correlation coefficient (P) of-0.48. From figure 5.17 it can be noted that the data points are spread away from the trend line showing a poor fit of Adj R^2 of 0.22. This trend can be attributed to the decrease in solar irradiance as the relative humidity increased hence leading to low PV power generation. This observation is in agreement with (Panjwani and Narejo, 2014).

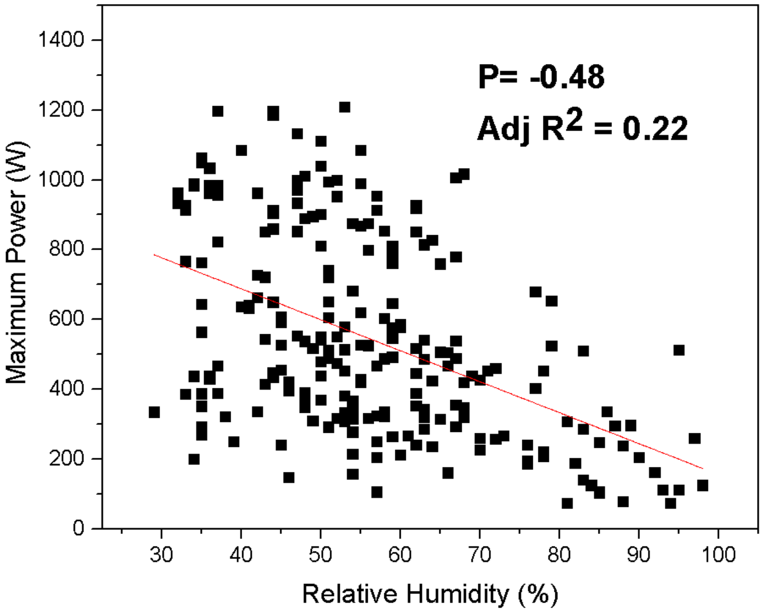


Figure 5.17: Variation of maximum power with relative humidity.

The relationship between the maximum power and relative humidity is expressed using equation 5.14 obtained from a linear fit with Adj R^2 of 0.22 indicating a weak linear relationship

$$P_m = -8.87(Hum) + 1042.8 \tag{5.14}$$

5.4.2: Effect of relative humidity on the PV system efficiency

Figure 5.18 shows the variation of PV system efficiency with relative humidity. The PV system efficiency was highest at relative humidity of 98%. It was noted that PV system efficiency varied linearly with relative humidity with positive correlation coefficient of 0.70. However, from figure 5.18 the data points are spread indiscriminately away from the line of fit indicating a weak relationship hence poor linear fit of Adj R^2 of 0.49.

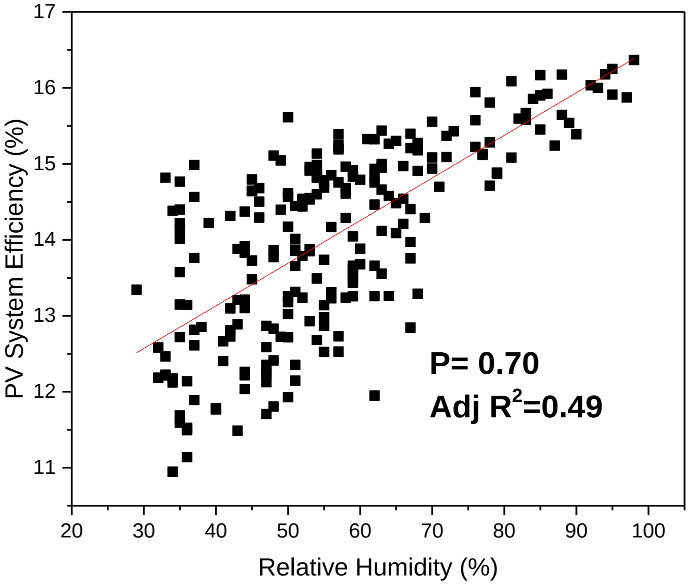


Figure 5.18: Variation of PV system efficiency with relative humidity.

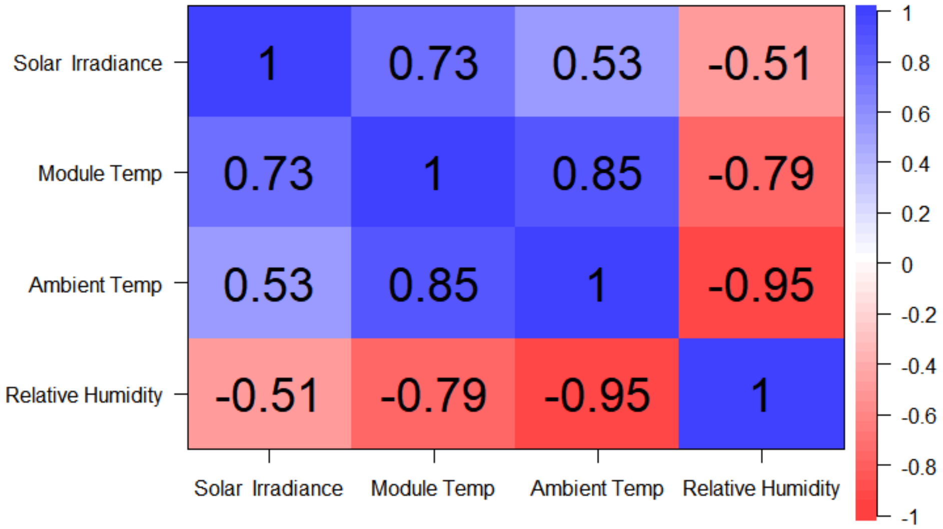
The relationship between the PV system efficiency (η) and relative humidity (Hum) is expressed by equation 5.15

$$\eta = 0.056(Hum) + 10.89 \tag{5.15}$$

5.5: Principal Component Analysis

Table 5.2 shows a correlation matrix containing Pearson's coefficient correlation values between measured weather parameters.

Table 5.2: Correlation matrix showing correlation between the measured weather parameters.

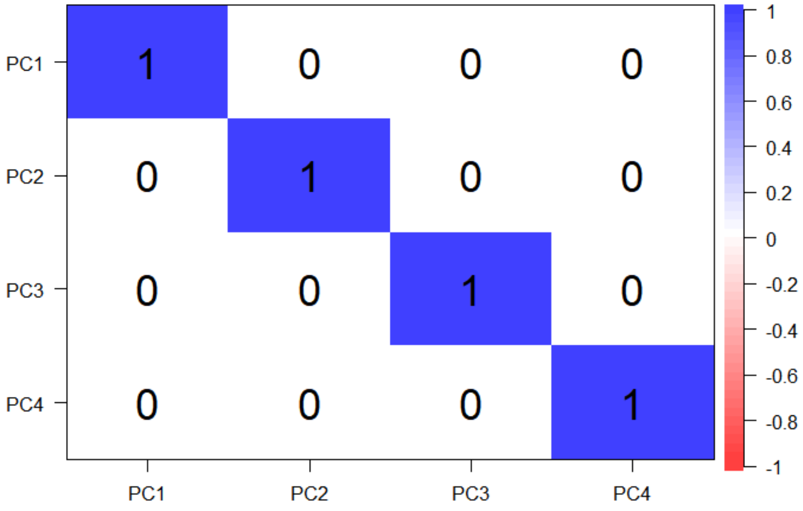


	Solar Irradiance	Module Temp	Ambient Temp	Relative Humidity
Solar Irradiance	1	0.73	0.53	-0.51
Module Temp	0.73	1	0.85	-0.79
Ambient Temp	0.53	0.85	1	-0.95
Relative Humidity	-0.51	-0.79	-0.95	1

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

The correlation coefficients shown in Table 5.2 greater than 0.70 indicate high correlation between variables hence indicate redundant information in the data which often decrease accuracy of a predictive model. Principal component analysis was used to remove redundant information from the measured weather parameters resulting into four uncorrelated principal components (PC's) as shown in Table 5.3.

Table 5.3: Correlation matrix showing correlation between the four principal components



	PC1	PC2	PC3	PC4
PC1	1	0	0	0
PC2	0	1	0	0
PC3	0	0	1	0
PC4	0	0	0	1

1
0.8
0.6
0.4
0.2
0
-0.2
-0.4
-0.6
-0.8
-1

The first (PC1), second (PC2), third (PC3) and fourth principal (PC4) components explained 79.86%, 15.53%, 3.70% and 1.113% of variance of the data respectively. The PCA results indicate that the first two principal components (PC1 and PC2) account for majority (95.19 %) of the variability of the dataset. Figure 5.19 shows resulting score plots and loading plots obtained from the PCA analysis. The score plots show distinct clustering of the data into three clusters based on the PV power output split into levels namely high, medium and low PV power output. Figure 5.19 also shows four vector lines namely for the four weather variables. The vector lines represent loadings plots showing how strongly the variables impacts PC1 and PC2. From Figure 5.19 it can be noted that the first principal component one (PC1) has a large positive association with irradiance, module temperature and ambient temperature while relative humidity has positive association with the second principal component (PC2). The angle between irradiance and module temperature show positive correlation between the two variables. The module temperature, ambient temperature and irradiance are negatively correlated with relative humidity. The loadings plot shows that solar irradiance is the furthest from the PC's origin hence has the most influence on PC1 making it the most important variable. This is in agreement with Kim *et al.*, (2019) reported that solar irradiance was most important variable with 43.7% importance.

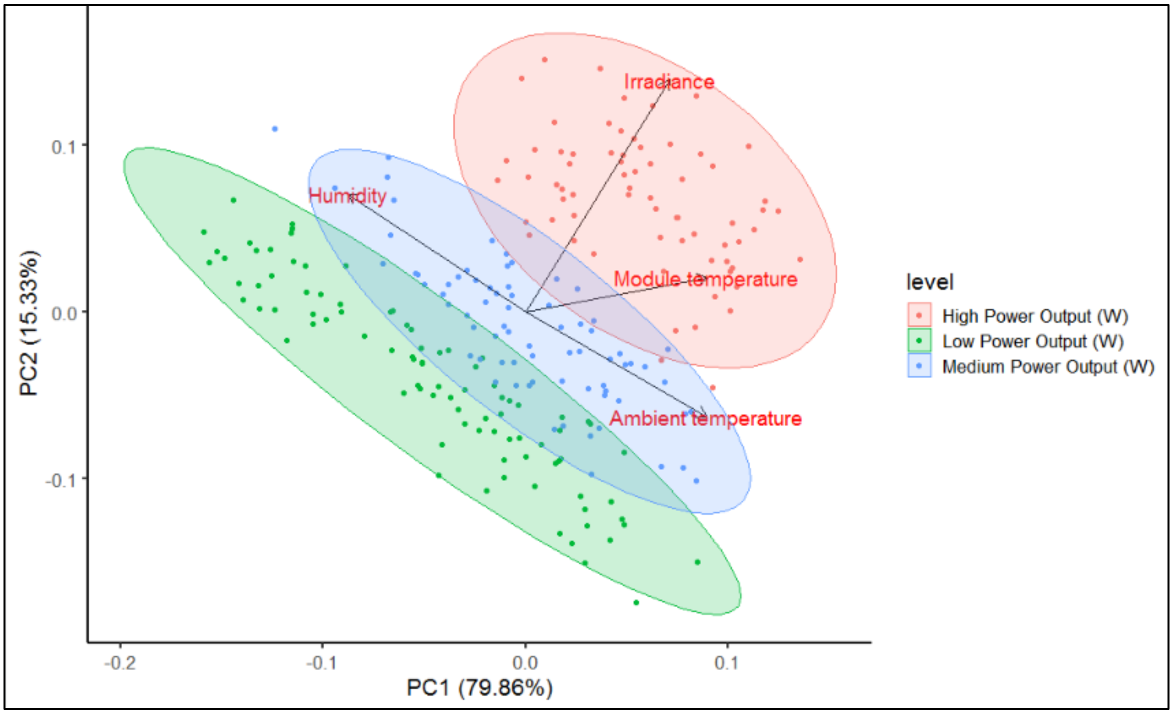


Figure 5.19: PCA biplot showing score plot and loadings plot between the three clusters of power; low, medium and high-power output.

5.6: Random Forest Regression

Table 5.4 shows the performance evaluation of random forest regression model based on *k-fold*, "LOOCV", random resampling cross validation techniques employed. Performance metrics were used to evaluate the model as shown in Table 5.4. The LOOCV yielded the optimum model with highest R^2 of 0.96 and the lowest MAE and RMSE of 51 and 65 on training dataset respectively compared to *k –fold* and CV random sampling.

Table 5.4: Performance evaluation of RFR training data set and test dataset based on k-fold, “LOOCV” and CV (Random resampling) employed

	Root Mean Absolute Error (RMSE)		**Coefficient of determination(R^2)**		**Mean Absolute Error (MAE)**	
Cross Validation technique	Train	Test	Train	Test	Train	Test
$\boldsymbol{k-fold(3)}$	76W	84.4	0.94	0.90	58.1	62.2
LOOCV	65W	94	0.96	0.87	51.8	68
CV (Random resampling)	79W	88.89	0.94	0.88	63.98	64.5

Figure 5.20 shows regression plot obtained from RFR model employing the “LOOCV” technique on the training dataset. The regression plot shows high correlation between the measured power output and the predicted power output with correlation coefficient (P) of 0.98. The Adj R^2 of 0.95 shows a good fit of the RFR model on training dataset.

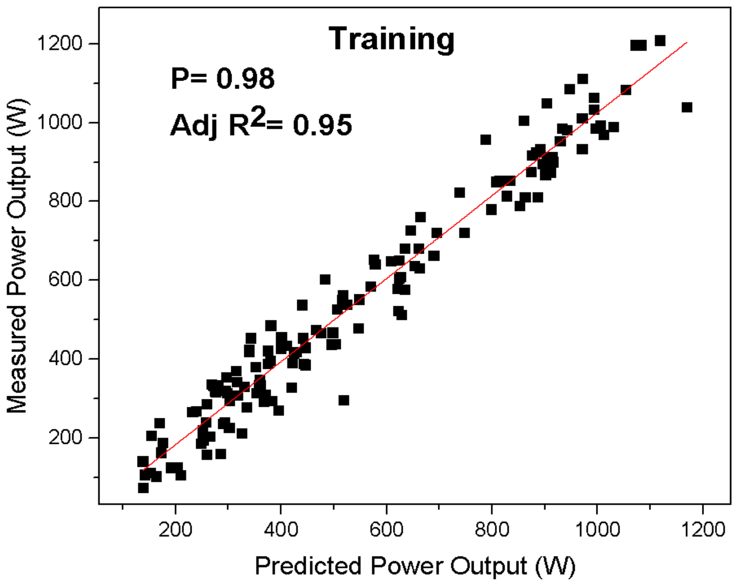


Figure 5.20: Measured power output against predicted power output by Random forest regression (RFR) model for the training dataset.

The trained model was tested using the testing data set. Figure 5.21 shows the regression plot based on the RFR model employing the "LOOCV" technique on the testing dataset. The testing dataset consisted of 44 observables and 5 variables. Figure 5.21 shows high correlation with correlation coefficient (P) of 0.93. The Adj R^2 of 0.87 shows a good fit of the RFR model on testing dataset.

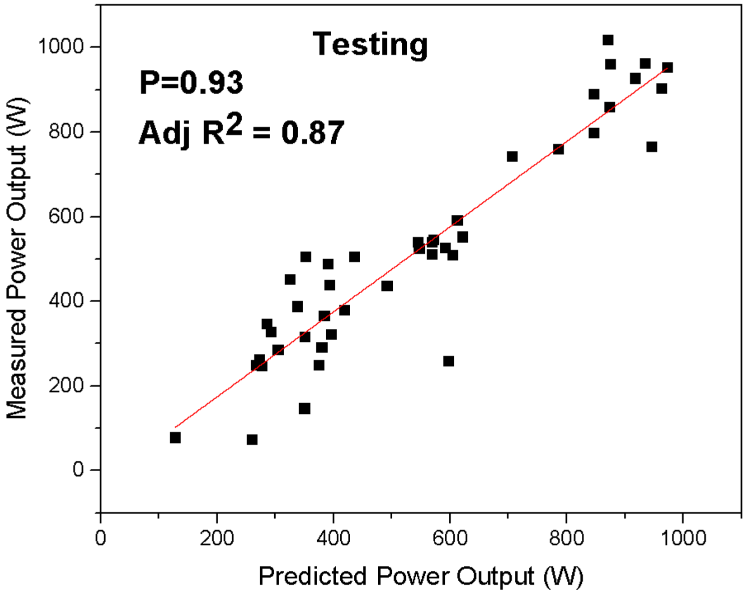


Figure 5.21: Measured power output against predicted power output by RFR model for the testing dataset.

5.7: Support Vector Regression

Table 5.5 shows performance evaluation of support vector regression model based *k -fold*, “LOOCV”, random resampling cross validation techniques employed. The performance of the trained model was evaluated. Table 5.5 shows LOOCV cross validation yielded the best trained model with R^2 of 0.98 and the lowest MAE and RMSE of 29.01W and 40.4W respectively compared to *k -fold* and cross validation random resampling techniques.

Table 5.5: Performance evaluation of SVR training data set and test dataset based on k-fold, "LOOCV" and CV (Random resampling) employed.

		Root mean square error (RMSE)		**Coefficient of determination (R^2)**		**Mean absolute error (MAE)**	
Cross Validation technique	Parameters	Train	Test	Train	Test	Train	Test
$\boldsymbol{k-fold(3)}$	C=1, sigma=0.05	40.70	45.10	0.98	0.97	30.40	29.01
LOOCV	C=1, sigma=0.05	40.40	45.10	0.98	0.97	29.01	29.27
CV (Random resampling)	C=45.12, sigma=0.12	47.96	50.30	0.97	0.96	31.72	32.06

Figure 5.22 shows the regression plot of SVR model employing the "LOOCV" technique on the training dataset. The regression plot shows high correlation with correlation coefficient (P) of 0.99. The Adj R^2 of 0.98 shows a good fit of the SVR model on training dataset.

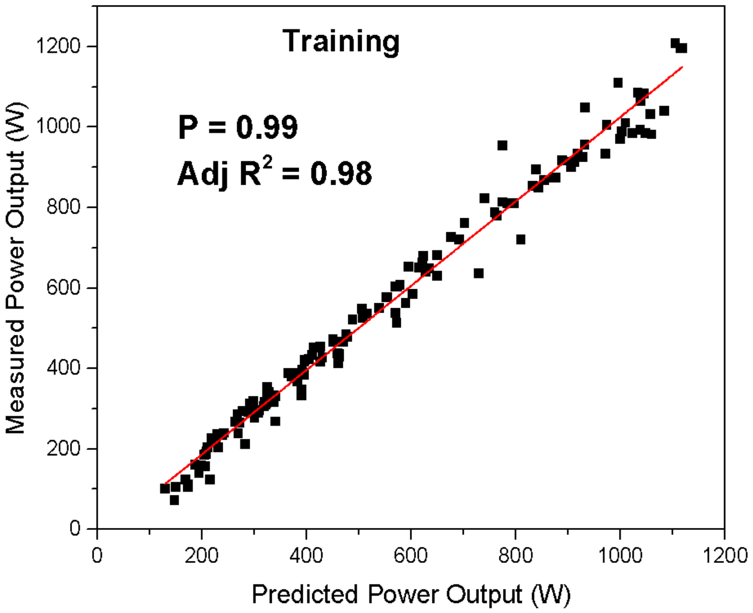


Figure 5.22: Measured power output against predicted power by Support vector regression (SVR) model for the training dataset.

Figure 5.23 shows regression plot for SVR model employing the "LOOCV" technique on the testing dataset. The regression plot shows high correlation with correlation coefficient (P) of 0.98. The Adj R^2 of 0.97 shows a good fit of the trained SVR model on the test dataset.

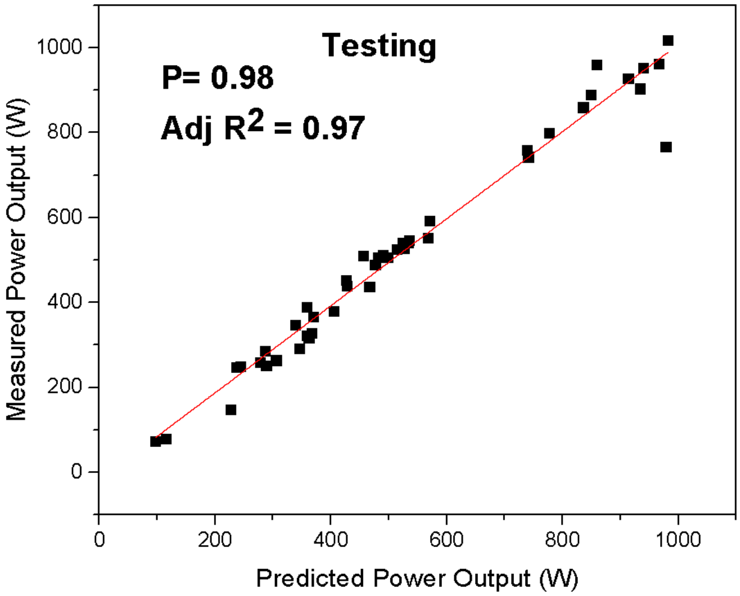


Figure 5.23: Measured power output against predicted power output by SVR for the testing dataset.

5.8: Model Validation

Model validation is the analysis of the goodness of a regression by checking the performance of the model using the validation dataset not used in model estimation is applied. The validation dataset consisting of 46 observables and 5 variables. Table 5.6 shows the comparison of performance of RFR and SVR trained model on validation dataset using RMSE, Adj R^2 and MAE accuracy metrics. The SVR performed better than the RFR yielding Adj R^2 of 0.97 and RMSE and MAE of 43.16 and 32.57 respectively on validation dataset. The RMSE of SVR model on validation dataset is 43.16 which is almost half of the RMSE of RFR model which is 86. This shows that SVR model is more robust than RFR and has capabilities of reducing errors during computations.

Table 5.6: Comparison of performance of RFR and SVR based on performance on validation dataset.

ML technique	RMSE		Adj R^2		MAE	
	Train	**Valid**	**Train**	**Valid**	**Train**	**Valid**
RFR	65	86	0.95	0.90	51.8	69
SVR	40.4	43.16	0.98	0.97	29.01	32.57

Figure 5.24 shows the measured power output against the predicted power output by RFR model employing the "LOOCV" technique on the validation dataset. The regression plot shows high correlation between the measured power output and the predicted power output with correlation coefficient (P) of 0.95 and Adj R^2 of 0.90.

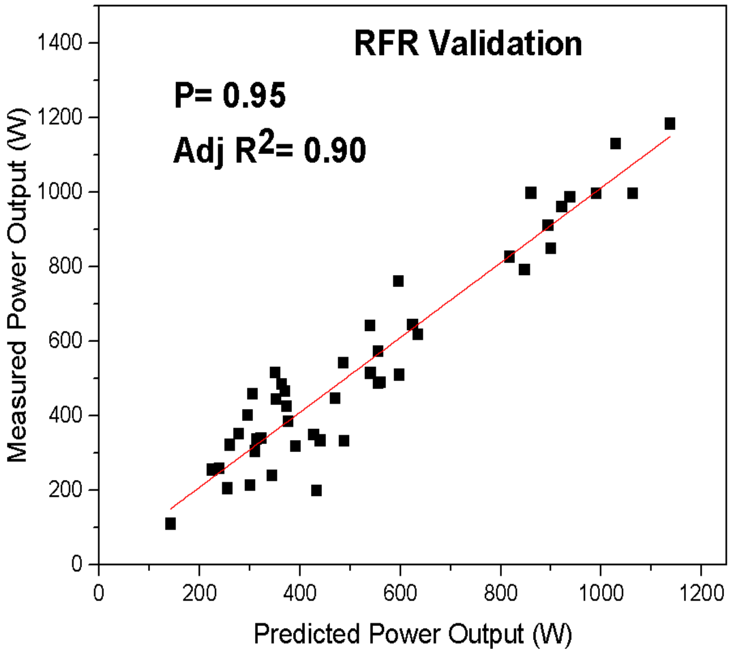


Figure 5.24: Measured power output against predicted power output by RFR model for validation dataset.

Figure 5.25 shows the measured power output against the predicted power output by SVR model employing the "LOOCV" technique on the validation dataset. The regression plot shows high correlation between the measured power output and the predicted power output with correlation coefficient (P) of 0.99 and Adj R^2 of 0.97 showing good fit of the model.

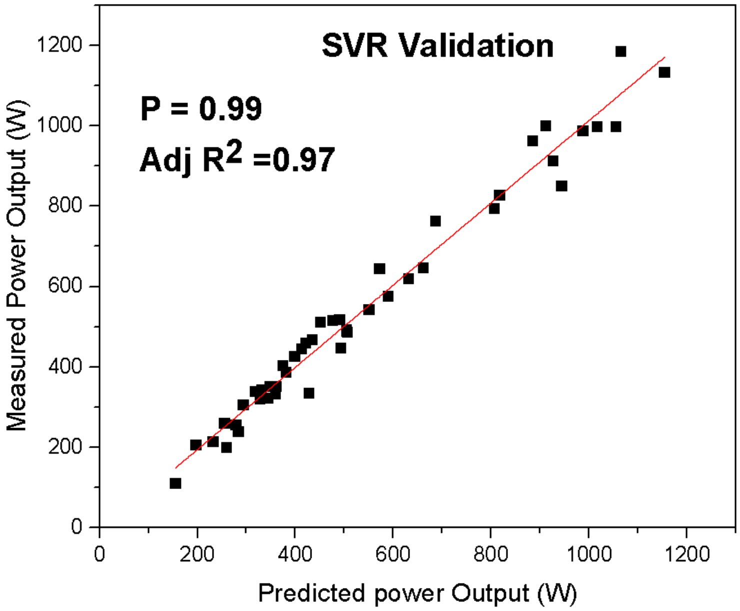
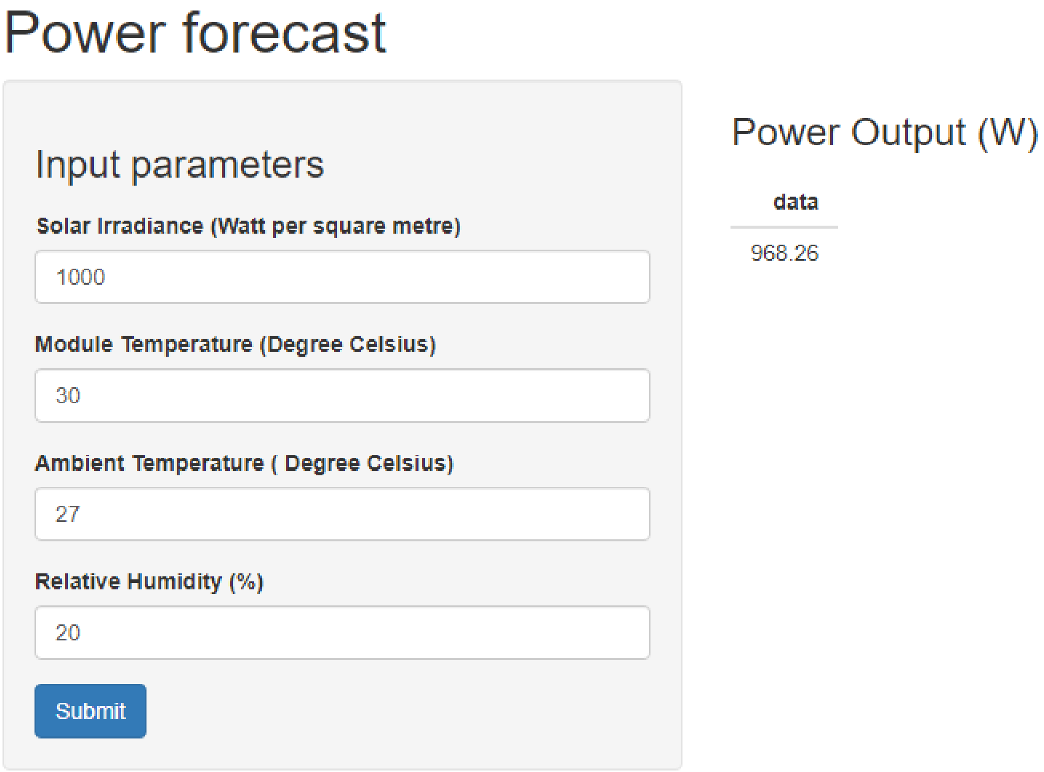


Figure 5.25: Measured power output against predicted power output by SVR model for the validation dataset.

5.9: Power forecast application based on the 1.5kW PV Solar system

The power forecast application was successfully built using the Shiny application in R environment based on the SVR model using the “LOOCV” technique. The application is equipped with input buttons namely; solar irradiance, module temperature, ambient temperature and relative humidity and then outputs the predicted PV power output based on the trained SVR model as shown in Figure 5.26.



Power forecast

Input parameters

Solar Irradiance (Watt per square metre)

1000

Module Temperature (Degree Celsius)

30

Ambient Temperature (Degree Celsius)

27

Relative Humidity (%)

20

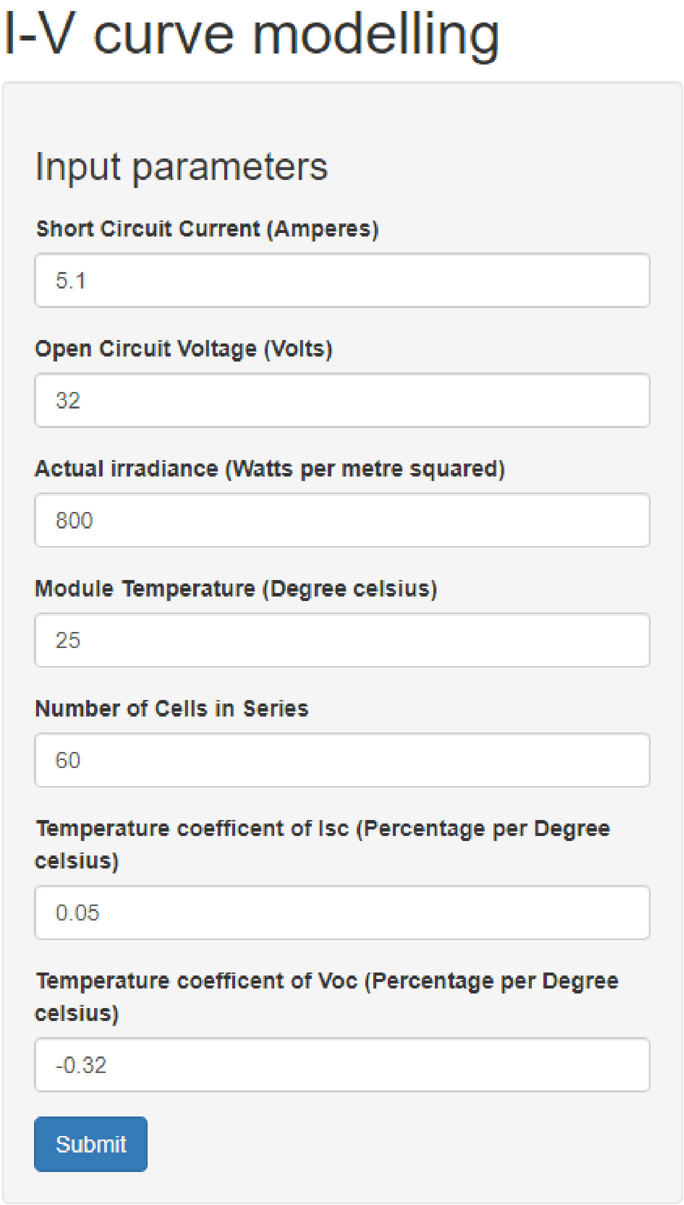
Power Output (W)

data
968.26

Figure 5.26: Power Forecast application user interface with input buttons, submit button and PV Output power.

5.10: I-V curve application

The I-V curve was modelled based on the single diode model in the Shiny app environment in R. The user interface enables one to input parameters, was designed as shown in the Figure 5.27.



I-V curve modelling

Input parameters

Short Circuit Current (Amperes)

5.1

Open Circuit Voltage (Volts)

32

Actual irradiance (Watts per metre squared)

800

Module Temperature (Degree celsius)

25

Number of Cells in Series

60

Temperature coefficent of Isc (Percentage per Degree celsius)

0.05

Temperature coefficent of Voc (Percentage per Degree celsius)

-0.32

Figure 5.27: I-V curve shinny app user interface display showing the input parameters numeric input buttons and submit button for execution.

Figure 5.28 shows corresponding I-V and P-V curves of the solar panel resulting from Figure 5.27.

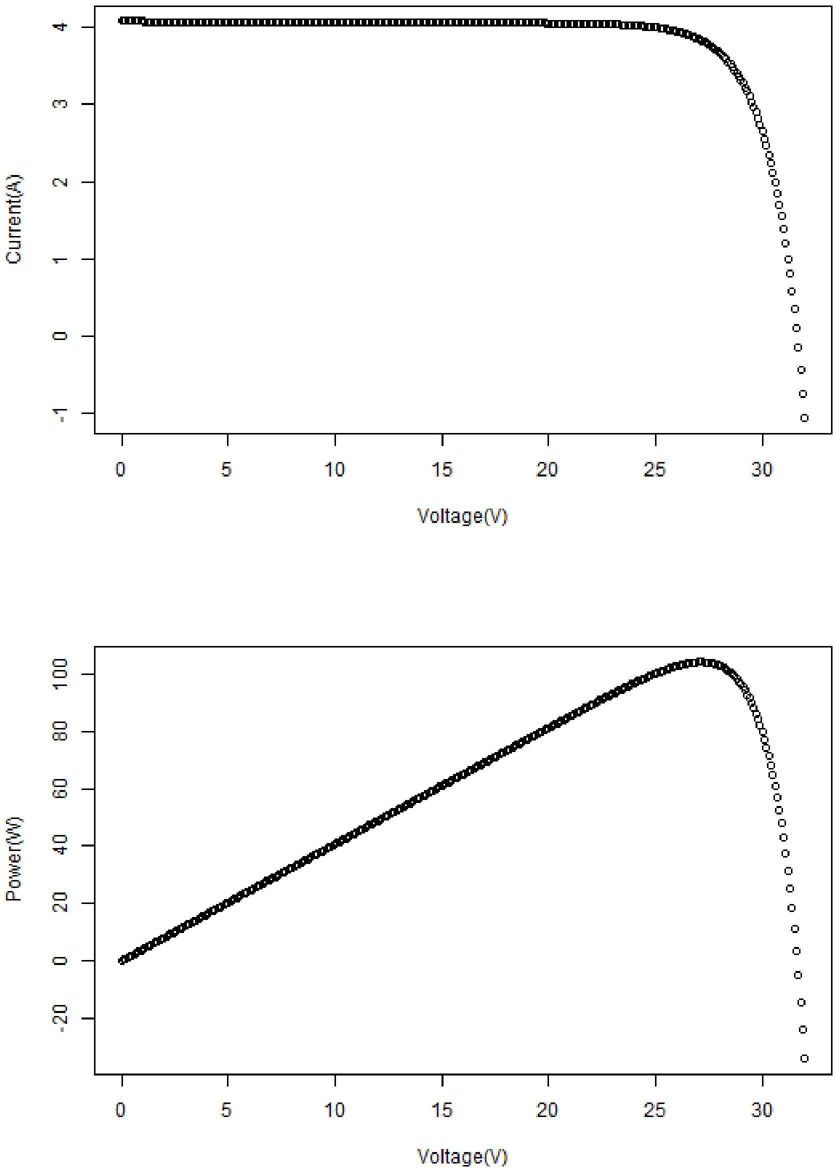


Figure 5.28: Current-Voltage, Power-Voltage plots image from the I-V curve application

The effect of irradiance on the performance of 250 W PV module was simulated using the I-V shiny application. The irradiance was varied from 1000W/m^2 to 200W/m^2 in steps of 200W/m^2. Figure 5.29 shows that increasing solar irradiance increases the current when the temperature is kept at a constant of 25 ℃.

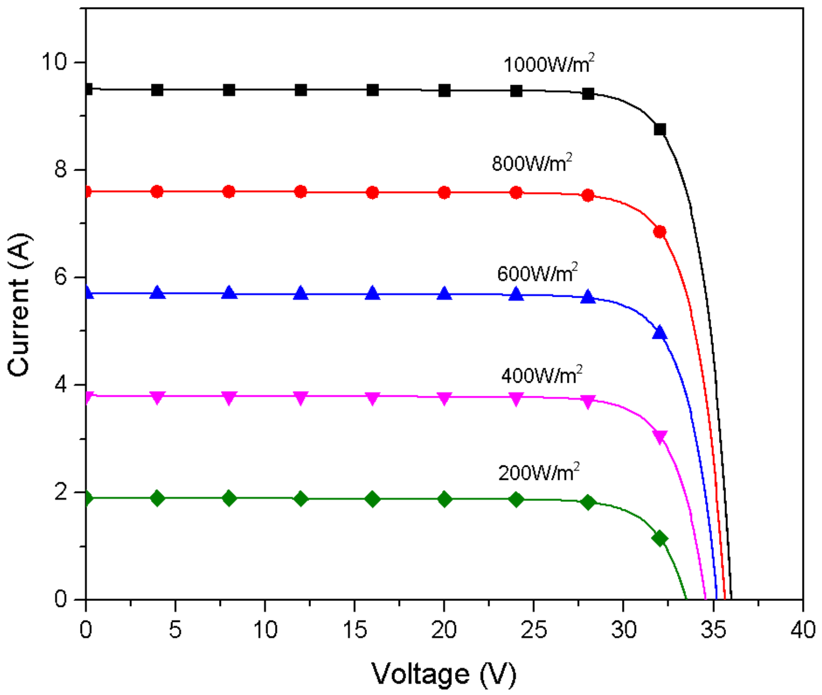


Figure 5.29: I-V curves for varying irradiance at a constant module temperature of 25℃

The effect of the module temperature on the performance of 250 W solar panel was simulated using the I-V shiny application. The module temperature was simulated from 25 ℃ to 85 ℃. Figure 5.30 shows that increasing module temperature decreases the open circuit voltage when the irradiance was kept at a constant of 1000W/m^2.

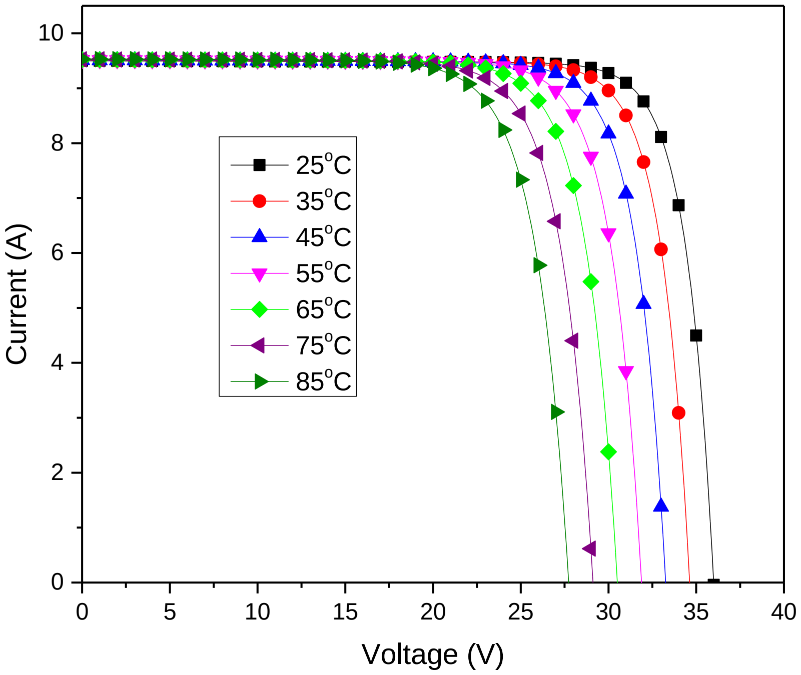


Figure 5.30: I-V curve at varying module temperature with a constant irradiance of 1000W/m^2

The performance of the I-V application was further examined by comparing the simulated I-V curve of the 1.5kW PV system with the experimental data obtained when the solar irradiance is 1117.5W/m^2 and module temperature of 42.8℃ as shown by the Figure 5.31. When compared to the experimental results, the simulated I-V curve has a higher short circuit voltage and open circuit voltage. This was attributed to the fact that the single diode equation only implements the impact of module temperature and solar irradiance on the performance of a module assuming the effect of dust, relative humidity on performance of modules.

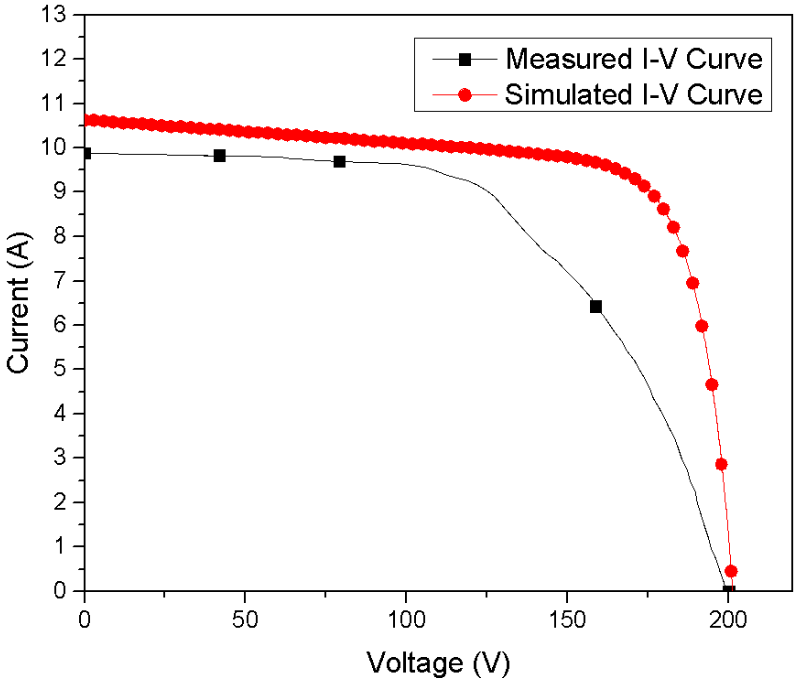


Figure 5.31: I-V curves of the simulated and measured at a solar irradiance 1117.5 W/m^2 and at a module temperature 42.8°C.

CHAPTER SIX: CONCLUSIONS AND RECOMMENDATIONS

6.1: Conclusions

An experimental setup was successfully installed to collect performance data of the 1.5kW PV system in varying weather conditions for a period of three weeks. Real-time data of solar irradiance, relative humidity, module temperature and ambient temperature was measured and studied. The study revealed that ambient temperature had a weak positive linear relationship with measured solar irradiance correlation coefficient (P) of 0.53 and Adj R^2 of 0.27 while the relative humidity showed a weak negative relation with solar irradiance with P of -0.51 and Adj R^2 of 0.25. The results also demonstrated that the ambient temperature and relative humidity are inversely proportional showing a strong negative correlation of -0.95 and Adj R^2 of 0.90. The ambient temperature and module temperature were found to be positively correlated with a correlation coefficient (P) of 0.85 and Adj R^2 of 0.72.

Real-time performance data of the 1.5kW PV system was collected and studied. The study focused on the influence of solar irradiance, ambient temperature and relative humidity on performance of photovoltaic system. From the results it was observed that solar irradiance had a very strong positive linear relationship with the power output generated by the PV system with positive correlation of 0.99 and Adj R^2 of 0.98. The short circuit current was also found to increase with increasing solar irradiance with high positive correlation P of 0.99 and Adj R^2 of 0.98. However, the series resistance and shunt resistance decreased with increasing solar irradiance forming exponential decay curve. The module efficiency decreased with the increasing amount of irradiance falling on the panels. It was further noted that open circuit voltage decreased with increasing module temperature indicating a very weak negative relationship with P of -0.50 and Adj R^2 of 0.25. Module efficiency exhibited a strong negative relationship module temperature with P of -0.87 and Adj R^2 of 0.76. The relative humidity increased with decreasing PV power output showing very weak relationship with P of -0.48 and Adj R^2 of 0.22 while module efficiency increased with increasing relative humidity P of 0.70 and Adj R^2 of 0.49.

PV power forecasting models coupled with PCA were built using SVR and RFR and were successfully trained, validated and tested to forecast real-time PV power output. SVR model employing LOOCV yielded the best model compared to *k -fold* and CV (Random resampling) cross validation techniques with RMSE of 40.4, Adj R^2 of 0.98 and MAE of 29.01 on training dataset and RMSE of 45.10, Adj R^2. of 0.97 and MAE of 29.27 on testing dataset. RFR model employing LOOCV yielded best model *k-fold* and CV (Random resampling) cross validation techniques with RMSE of 65, Adj R^2 of 0.95 and MAE of 51.8 on training dataset whereas for testing set RMSE of 94, AdjR^2 of 0.87, MAE of 68 were obtained. The trained models were further evaluated using validation dataset, SVR model outperformed RFR with RMSE of 43.16, Adj R^2 of 0.97 and MAE of 32.57 compared to RMSE of 86, Adj R^2 of 0.90 and MAE of 69 obtained from RFR model. The SVR model based shiny application to predict 1.5kW Solar PV power output when weather parameters inputs were built. I-V and P-V model based on single diode model interactive Shiny application was built successfully using in R Studio. The I-V curve model demonstrated the influence of varying irradiance and module temperature on performance of solar modules.

6.2: Recommendation for further work

Data collection for a longer period is recommended to ensure a large dataset is used for both training and testing hence increasing model accuracy. Collection of data from different seasons is recommended to ensure accurate prediction of PV power over different seasons.

To ensure applicability of the predictive model to different PV solar systems at different locations, it is recommended that data be collected from different sites with varying weather conditions to ensure inclusivity in the interactive predictive models built.

REFERENCES

Abuella, M. and Badrul, C. (2016). Solar Power Forecasting Using Support Vector Regression. in *Proceedings of the American Society for Engineering Management*.1-6

Achieng, P. F., Davidsdottir, B., and Birgir, I. (2012). Potential contribution of geothermal energy to climate change adaptation: A case study of the arid and semi-arid eastern Baringo lowlands, Kenya. *Renew. Sustain. Energy Rev.* **16**, 4222–4246.

Africa Energy Series, I. (2020). Kenya Special report.

Al-Dahidi, S., Osama, A., Jehad, A., Mohammad, A., and Bashar, R.S. (2018). Extreme Learning Machines for Solar Photovoltaic Power Predictions. *energies.***11**, 2725.

Alanazi, M., Alanazi, A., and Khodaei, A. (2017). Long-Term Solar Generation Forecasting *IEEE/PES Transmission and Distribution Conference and Exposition (T&D)* ,1-5.

Aliberti, A., Bottaccioli, L., Cirrincione, G., Macii, E., Acquaviva, A., and Patti, E. (2018). Forecasting short-term solar radiation for photovoltaic energy predictions. *SMARTGREENS - Proc. 7th Int. Conf. Smart Cities Green ICT Syst.*, 44–52.

Almohri, H., Du, C., Hu, Z., and Wang, J. (2014). Data Analysis and Prediction of Power Generated by Photovoltaic Systems. *WCECS*. **II**, 22–24.

Arjyadhara, P. and Chitralekha, J. (2013). Analysis of Solar PV cell Performance with Changing Irradiance and Temperature. *Int. J. Eng. Comput. Sci.* **2**, 214–220.

Bashir, M. A., Ali, H. M., Khalil, S., Ali, M., and Siddiqui, A. M. (2014). Comparison of Performance Measurements of Photovoltaic Modules during Winter Months in Taxila, Pakistan. *Intl Journal of Photoenergy,,* **2014**,1-8.

Behrens, J. T. (1997). Principles and Procedures of Exploratory Data Analysis. **2**, 131–160.

Benda, V. and Machacek, Z. (2016). A note on parameters of photovoltaic cells in dependence on irradiance and temperature.

Bhattacharya, T., Chakraborty, A. K., and Pal, K. (2014). Effects of Ambient Temperature and Wind Speed on Performance of Monocrystalline Solar Photovoltaic Module in Tripura, India, *Journal of Solar Energy* ,1-6.

Breiman, L. (2001). Random Forests. University of California Berkeley, CA 94720 .1–33.

Cervone, G., Clemente-Harding, L., Alessandrini, S., and Delle Monache, L. (2017). Short-term photovoltaic power forecasting using Artificial Neural Networks and an Analog Ensemble. *Renew. Energy* **108**, 274–286.

Ciulla, G., Brano, V. Lo, and Moreci, E. (2013). Forecasting the cell temperature of PV modules with an adaptive system. *Int. J. Photoenergy* **2013,** 1-11.

Dhar, S., Scholar, P. G., Sridhar, R., G, A. S., Avasthy, V., and Scholar, P. G. (2010). Modeling and Simulation of Photovoltaic Arrays, *17th National Power Systems Conference.1*–5.

Diantoro, M., Suprayogi, T., Hidayat, A., Taufiq, A., Fuad, A., and Suryana, R. (2018). Shockley ' s Equation Fit Analyses for Solar Cell Parameters from I-V Curves. *Int. J. Photoenergy* **2018**, 1-8.

Elkholy, A. and El-ela, A. A. A. (2019). Optimal parameters estimation and modelling of photovoltaic modules using analytical method. *Heliyon* **5**, e02137.

Giorgi, M. G. De, Congedo, P. M., and Malvoni, M. (2014). Photovoltaic power forecasting using statistical methods : impact of weather data. *IET Science, Measurement and Technology,***8,** 90-97.

Goswami, D. Y. (2017). Solar energy resources. *Energy Conversion,**2**.*, 85–136.

Ho, S. M. (2019). Principal Components Analysis (PCA). University of Georgia, Athens, GA 30602-2501

IES (2019). Environmental Impact of Fossil Fuel. 4964706.

Isaksson, E. and Conde, M. K. (2018). Solar Power Forecasting with Machine Learning Techniques. **46**, 1-64.

Italia, HT Instruments. (2012). HT304N Manual.

Jang, H. S., Kuk, B., Hong, P., and Dan, K., (2016). Solar Power Prediction Based on Satellite Images and Support Vector Machine, 1-10

Diantoro, M., Suprayogi, T., Hidayat, A., Taufiq, A., Fuad, A., and Suryana, R. (2018). Shockley ' s

Equation Fit Analyses for Solar Cell Parameters from I-V Curves . *Intl J. of Photoenergy* **2018,**1-7.

Kanwal, S. (2018). Support Vector Machine and Gaussian Process Regression based Modeling for Photovoltaic Power Prediction. *2018 Int. Conf. Front. Inf. Technol.*, 117–122.

Kassambara, A. (2011). Practical Guide to Principal Component in R. 1st edition

Khandakar, A., Chowdhury, M. E. H., Kazi, M.-K., Benhmed, K., Touati, F., Al-hitmi, M., and Gonzales, A. J. S. P. (2019). Machine Learning Based Photovoltaics (PV) Power Prediction Using Different Environmental Parameters of Qatar. *Energies, MDPI, Open Access Journal,* **12**,1-19.

Kim, S.-G., Jung, J.-Y., and Sim, M. (2019). A Two-Step Approach to Solar Power Generation Prediction Based on Weather Data Using Machine Learning. *Sustainability* **11**, 1501.

Kostylev, V. and Pavlovski, A. (2011). Solar power forecasting performance - towards industry standards. *1st Int. Work. Integr. Sol. Power into Power Syst. Aarhus, Denmark*, 1–8.

Lenzmann, F. and Carol. O. (2016). The social and economic consequences of the fossil fuel supply chain. *MRS Energy & Sustainability*,1–32.

Madhiarasan, M. and Deepa, S. N. (2017). Review of Forecasters Application to Solar Irradiance Forecasting. *IJSRCSEIT*, **2**, 26–30.

Mark, H. Anjali ,S. and Paul K.(2009). Target Market analysis: Kenya's Solar Energy Market.*gtz* 9–13.

Mekhilef, S., Saidur, R., and Kamalisarvestani, M. (2012). Effect of dust , humidity and air velocity on efficiency of photovoltaic cells. *Renew. Sustain. Energy Rev.* **16**, 2920–2925.

Ministry of Energy (2012). Feed-In-Tarrifs Policy on Wind, Biomass, Small-Hydro , Geothermal , Biogas and Solar Resource Generated Electricity Initial Issue : March 2008. 1–17.

Musanga, L. M., Barasa, W. H., and Maxwell, M. (2018). The Effect of Irradiance and Temperature on the Performance of Monocrystalline Silicon Solar Module in Kakamega. *Phys. Sci. Int. J.* **19**, 1–9.

Nicholas N. Tasie, Israel-Cookey, and Banyie, L. J. (2018). The Effect of Relative Humidity on the

Solar Radiation Intensity in Port Harcourt, Nigeria. *Int. J. Res. Eng. Technol.* **5**, 9.

Omubo-Pepple, V. B., Israel-Cookey, C., and Alaminokuma, G. I. (2009). Effects of temperature, solar flux and relative humidity on the efficient conversion of solar energy to electricity. *Eur. J. Sci. Res.* **35**, 173–180.

Panjwani, M. K. and Narejo, G. B. (2014). Effect of Humidity on the Efficiency of Solar Cell (photovoltaic).*iJERGS.***2**, 499–503.

Paulin, B. J. and Praynlin, E. (2016). Solar photovoltaic output power forecasting using back propagation neural network. *ijsc.* 1144-1152.

Pearsons, R. (2018)., Exploratory data analysis: Second Look, In: Vipin Kumar, *Exploratory Data Analysis Using R* "CRC press." 409-422.

Rahman, M. A., Hossain ,M., and Jakaria, A.H., (2018). Smart Weather Forecasting Using Machine Learning : A Case Study in Tennessee Smart Weather Forecasting Using Machine Learning : A Case Study in Tennessee. *ArXiv, abs/2008.10789*.

Salcedo-Sanz, S., Casanova-Mateo, C., Pastor-Sánchez, A., and Sánchez-Girón, M. (2014). Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization - Extreme Learning Machine approach. *Sol. Energy* **105**, 91–98.

Samoita, D., Nzila, C., and Østergaard, P. A. (2020). Barriers and Solutions for Increasing the Integration of Solar Photovoltaic in Kenya' s Electricity Mix. *Energies* **13**, 5502.

Schmidt, T., Calais, M., Roy, E., Burton, A., Heinemann, D., Kilper, T., and Carter, C. (2017). Short-term solar forecasting based on sky images to enable higher PV generation in remote electricity networks. *Renew. Energy Environ. Sustain*. **23**, 0–5.

Shrestha, A. K., Thapa, A., and Gautam, H. (2019). Solar Radiation , Air Temperature , Relative Humidity , and Dew Point Study : Damak , Jhapa , Nepal. *Intl J. of Photoenergy.* 1-8.

Smola, A. J. and Olkopf, B. S. C. H. (2004). A tutorial on support vector regression. *Statistics and Computing* **14**, 199–222.

Solargis (2019). Kenya_PVOUT_mid-size-map_156x220mm-300dpi_v20191015.

Solinic EA.Ltd (2017). Grid Tie Solar Modules 250/300Wp. *In* "Grid Tie Solar Modules 250/300Wp

East Africa Ltd.”

IEA, (2019). U. S. Department of Energy ,International Energy Outlook 2019.

Vanderplas, J. (2016). Machine Learning.In: Dawn S., “*Python Data Science Handbook.*” O’Reilly, CA, **1** ,334-346.

Wan, C., Jian .Z, Yonghua, S. Zhao, X., Jin, L. and Zechun, H. (2015). Photovoltaic and Solar Power Forecasting for Smart Grid Energy Management. *CSEE Journal of power and energy systems*. **1**, 38–46.

Whitaker, C. M., Townsend, T. U., Wenger, H. J., Iliceto, A., Chimento, G., and Paletta, F. (1992). Effects of irradiance and other factors on PV temperature coefficients. *Conf. Rec. IEEE Photovolt. Spec. Conf.* **1**, 608–613.

Sharma, N., Sharma, P., Irwin, D.E., and Shenoy, P.J. (2011). Predicting solar generation from weather forecasts using machine learning.*2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 528-533.

Yerrapragada, J. B., Bharath K. S., and Mayukh S.(2013). Short-Term Power Forecasting of Solar PV Systems Using Machine Learning Techniques.1–5.

Zafarani, R., Eftekharnejad, S., and Patel, U. (2018). Assessing the Utility of Weather Data for Photovoltaic Power Prediction. *ArXiv, abs/1802.03913*.

Zeman, M., Klaus, J., Olindo, I., Arno, H.M., and René A.C.M.M., (2014). Solar cell parameters and Equivalent circuit. *Solar Energy,* Delft University of Technology: Delft ,101.

APPENDICES

Appendix 1: Correlation and PCA script in R

```
library(corrplot)
library(psych)
library(ggfortify)
library(ggplot2)
library(ggpubr)
library(colorspace)
library(pastecs)
library(cluster)
library(dendextend)
theme_set (theme_classic (base_size = 16))
library(readily)
memo <- read excel("mumo.xlsx")
star(mumo)
summary(mumo)
corttt<-cor(mumo[,c(2:5)])
cor.plot(corttt)
 corrpp<-cor(mumo[1:5])
 cor.plot(corrpp)
mumo$level<-as.factor(mumo$level)
str(mumo)
##########################PCA
pca_res <- prcomp(mumo[2:5],scale. = TRUE)
autoplot(pca_res,data = mumo,colour='level',abline = TRUE)
autoplot(pca_res,data = mumo,colour='level',loadings = TRUE,
        loadings.colour='black', loadings.label=TRUE,
     loadings.label.size=5 ,frame.type='norm')
#####################################
summary(pca_res)
plot(pca_res,type="l")
pca_res$x
pca<-as.data.frame (pca_res$x)
corpca<-cor(pca)
cor.plot(corpca)
pca_res$rotation
library(dplyr)
newdata<- bind_cols(Pmax=mumo$Pmax,pca)
write.csv(newdata,"newdata.csv")
```

Appendix 2: SVM R script

```
library(caret)
library(plyr)
library(recipes)
library(dplyr)
library(doParallel)
library(keras)
data1 <- read.csv("newdata.csv")
str(data1)
#######################################################################
set.seed(12345)
rctrl1 <- trainControl(method = "cv", number = 3, returnResamp = "all",
                       savePredictions = "final")
rctrl2 <- trainControl(method = "LOOCV",savePredictions = "final")
rctrlR <- trainControl(method = "cv", number = 3, returnResamp = "all",
                       search = "random",savePredictions = "final")
rctrlAdapt <- trainControl(method = "adaptive_boot", number = 15,
                           search = "random",
                           predictionBounds = c(TRUE, FALSE),
                           savePredictions = "final")
# Data Partition
set.seed(12)
ind <- sample(3, nrow(data1), replace = TRUE, prob = c(0.6, 0.2,0.2))
training_data1 <- data1[ind==1,]
testing_data1 <- data1[ind==2,]
validation_data1<-data1[ind==3,]
trainX_data1 <- training_data1[,2:5]
```

```
trainY_data1 <- training_data1$Pmax
data1_rec_reg <- recipe(Pmax ~ ., data = training_data1) %>%
        step_center(all_predictors()) %>%
        step_scale(all_predictors())
testX_data1 <- testing_data1[,2:5]
testY_data1 <- testing_data1$Pmax
validX_data1 <- validation_data1[,2:5]
validY_data1<-validation_data1$Pmax
########################################################
set.seed(849)
SVMR_test_data1_reg_cv_model <- train(trainX_data1, trainY_data1,
                                       method = "svmRadial",
                                       tuneGrid = data.frame(.C = c(.25, .5, 1),
                                                             .sigma = .05),
                                       trControl = rctrl1,
                                       preProc = c("center", "scale"))
print(SVMR_test_data1_reg_cv_model)
plot(y = SVMR_test_data1_reg_cv_model$pred$obs, x
=SVMR_test_data1_reg_cv_model$pred$pred,
     xlab = 'prediction',
     ylab = 'observed')
abline(a = 0, b = 1, lty = 2, col = 2)
SVMR_test_data1_reg_pred <- predict(SVMR_test_data1_reg_cv_model, testX_data1)
postResample(pred = SVMR_test_data1_reg_pred, obs = testY_data1)
# plot observed vs predicted values for training and test set
plot(y = SVMR_test_data1_reg_pred, x = testY_data1,
     xlab = 'prediction',
     ylab = 'observed')
abline(a = 0, b = 1, lty = 2, col = 2)
########################################################################
set.seed(849)
SVMR_valid_data1_reg_cv_model <- train(trainX_data1, trainY_data1,
                              method = "svmRadial",
                              tuneGrid = data.frame(.C = c(.25, .5, 1),
                                               .sigma = .05),
                              trControl = rctrl1,
                              preProc = c("center", "scale"))
print(SVMR_valid_data1_reg_cv_model)
SVMR_valid_data1_reg_prad <- predict(SVMR_valid_data1_reg_cv_model, validX_data1)
plot(y =validY_data1, x =SVMR_valid_data1_reg_prad,
     xlab = 'prediction',
     ylab = 'observed')
abline(a = 0, b = 1, lty = 2, col = 2)
postResample(pred = SVMR_valid_data1_reg_prad, obs = validY_data1)
```

Appendix 3: Random forest R script

```
library(caret)
library(plyr)
library(recipes)
library(dplyr)
library(doParallel)
library(keras)
data1 <- read.csv("newdata.csv")
str(data1)
#ridomil Data Partition
set.seed(12)
ind <- sample(3, nrow(data1), replace = TRUE, prob = c(0.6, 0.2,0.2))
training_data1 <- data1[ind==1,]
testing_data1 <- data1[ind==2,]
validation_data1<-data1[ind==3,]
trainX_data1 <- training_data1[,2:5]
trainY_data1 <- training_data1$Pmax
data1_rec_reg <- recipe(Pmax ~ ., data = training_data1) %>%
  step_center(all_predictors()) %>%
  step_scale(all_predictors())
testX_data1 <- testing_data1[,2:5]
testY_data1 <- testing_data1$Pmax
validX_data1 <- validation_data1[,2:5]
validY_data1 <- validation_data1$Pmax
# random forest
set.seed(12)
seeds <- vector(mode = "list", length = nrow(training_data1) + 1)
seeds <- lapply(seeds, function(x) 1:20)
rctrl1 <- trainControl(method = "cv", number = 3, savePredictions = "final",
                       returnResamp = "all", seeds = seeds)
rctrl2 <- trainControl(method = "LOOCV", seeds = seeds,
                       savePredictions = "final")
rctrlR <- trainControl(method = "cv", number = 3,
                       returnResamp = "all", search = "random",
                       savePredictions = "final")
####################################################

set.seed(849)
RF_test_data1_reg_cv_model <- train(trainX_data1, trainY_data1,
                                    method = "rf",
                                    trControl = rctrl1,
                                    preProc = c("center", "scale"),
                                    ntree = 20,
                                    importance = TRUE)
print(RF_test_data1_reg_cv_model)
summary(RF_test_data1_reg_cv_model)

RF_test_data1_reg_cv_model1 <- predict(RF_test_data1_reg_cv_model,
                                       testX_data1)
postResample(pred = RF_test_data1_reg_cv_model1, obs = testY_data1)
# plot observed vs predicted values for training and test set

plot(y = RF_test_data1_reg_cv_model1, x = testY_data1,
     xlab = 'prediction',
    ylab = 'observed')
abline(a = 0, b = 1, lty = 2, col = 2)
#####################################################
set.seed(849)
RF_valid_data1_reg_cv_model <- train(trainX_data1, trainY_data1,
                             method = "rf",
                             trControl = rctrl1,
                             preProc = c("center", "scale"),
                             ntree = 20,
                             importance = TRUE)
print(RF_valid_data1_reg_cv_model)
summary(RF_valid_data1_reg_cv_model)

RF_valid_data1_reg_cv_model1 <- predict(RF_valid_data1_reg_cv_model,
                               validX_data1)
postResample(pred = RF_valid_data1_reg_cv_model1, obs = validY_data1)
# plot observed vs predicted values for training and test set

plot(y = RF_valid_data1_reg_cv_model1, x = validY_data1,
    xlab = 'prediction',
    ylab = 'observed')
abline(a = 0, b = 1, lty = 2, col = 2)
```

Appendix 4: I-V curve modelling

```
#import libraries
library(shiny)
library(shinythemes)
# User Interface                    #
ui <- pageWithSidebar(
  # Page header
  headerPanel('I-V curve modelling'),
  # Input values
  sidebarPanel(
    #HTML("<h3>Input parameters</h3>"),
    tags$label(h3('Input parameters')),
    numericInput("Iscn",
                 label = "Short Circuit Current (Amperes)",
                 value = "5.1"),
    numericInput("Vocn",
                 label = "Open Circuit Voltage (Volts)",
                 value = "32"),
    numericInput("G",
                 label = "Actual irradiance (Watts per metre squared)",
                 value = "800"),
    numericInput("Tn",
                 label = "Module Temperature (Degree celsius)",
          value = "25"),
  numericInput("Ns",
          label="Number of Cells in Series",
          value="60"),
  numericInput("Ki",
          label="Temperature coefficent of Isc (Percentage per Degree celsius)",
          value="0.05"),
  numericInput("Kv",
          label="Temperature coefficent of Voc (Percentage per Degree celsius)",
          value="-0.32"),
  actionButton("submitbutton", "Submit",
          class = "btn btn-primary")
),
mainPanel(
  # Output: plot ----
  plotOutput('plot1'),
  plotOutput('plot2'),
 # tags$label(h3('Maximum Power')),
  #textOutput('Pm'),
  textOutput("I"),
  textOutput("V"),
  textOutput("P")
)
```

```
)
# Server                    #
server <- function(input, output, session){
  # Input Data
  Vtn<-reactive({input$Ns*((1.38065e-23*(input$Tn+273)/1.602e-19)) })
# Equation 2
  I0n<-reactive({input$Iscn/((exp(input$Vocn/(1*Vt())))-1)  })                    # Equation 5
  I0<-reactive({I0n()*(((input$Tn+273)/298)^3)*exp((1.602e-19*1.2/(1*1.38065e-23))*(((1/298)-
1/(input$Tn+273)))) })        #Equation 4
  Ipvn<-reactive({input$Iscn})
  Ipv<-reactive({(input$G/1000)*(Ipvn()+(input$Ki/100)*((input$Tn+273)-298))})
#Equation 3
  Im<-reactive({input$Im(input$G/1000)})
  #Vm<-reactive({input$Vm-((input$Kv/100)*((input$Tn+273)-298))})
  #Pm<-reactive({input$Im*input$Vm}-((input$Kp)/100*((input$Tn+273)-298)))
  Vt<-reactive({input$Ns*((1.38065e-23*298)/1.602e-19)})
  Voltage<-reactive({ Voc<-input$Vocn+((input$Kv)/100*((input$Tn+273)-298))
  seq.default(0,Voc,0.1)})
  Ia<-reactive({seq.default(Ipvn(),0)})
  It1<-reactive({I0()*(exp((Voltage()+Ia()*0.001)/(Vtn()*1))-1)   })        #Part of Equation 1
  It2<- reactive({(Voltage()+Ia()*0.001)/1000 })                    # Part of Equation 1
  Current<-reactive({Ipv()-(It1()+It2())})
  Power<-reactive({Current()*Voltage()})
  # output
  output$plot1 <- renderPlot({
    plot(y = Current(), x = Voltage(),
              xlab = 'Voltage(V)',
              ylab = 'Current(A)')
  })
  output$plot2 <- renderPlot({
    plot(y = Power(), x = Voltage(),
         xlab = 'Voltage(V)',
         ylab = 'Power(W)')
  })
  output$I <-renderPrint({Current()
    write.csv(Current(),"Current.csv")
  })
  output$V <-renderPrint({Voltage()
    write.csv(Voltage(),"Voltage.csv")})
  output$P <-renderPrint({Power()
    write.csv(Power(),"Power.csv")})
}
shinyApp(ui = ui, server = server)
```

Appendix 5: Power shiny app

```
library(shiny)
library(data.table)
library(shinythemes)
library(caret)
library(plyr)
library(recipes)
library(dplyr)
library(doParallel)
library(keras)
setwd("D:/rita R/ndungu/Whita")
model <- readRDS("fodel.rds")
# User Interface #
ui <- pageWithSidebar(
  headerPanel('Power forecast'),
  sidebarPanel(
    #HTML("<h3>Input parameters</h3>"),
    tags$label(h3('Input parameters')),
    numericInput("Irradiance",
                 label = " Solar Irradiance (Watts per metre squared",
                 value = "1000"),
    numericInput("modtemp",

                 label = "Module Temperature (Degree Celsius",
                 value = "30"),
    numericInput("ambtemp",

                 label = "Ambient Temperature (Degree Celsius)",
                 value = "27"),
    numericInput("Humidity",
              label = "Relative Humidity (%)",
              value = "20"),

    actionButton("submitbutton", "Submit",
              class = "btn btn-primary")

  ),
  mainPanel(
    tags$label(h3('Power Output (W)')), # Status/Output Text Box
    verbatimTextOutput('contents'),
    tableOutput('Power')
  )
)
server <- function(input, output, session) {
  # Prediction results table
  # Input Data
  output$Power <- renderTable({
    test<- cbind(input$Irradiance,input$modtemp,input$ambtemp,input$Humidity)
    colnames(test)<-c("Irradiance","modtemp","ambtemp","Humidity")
    f<-write.csv(test,"test.csv")
    p<- read.csv("test.csv")
    p<-p[,-1]
    Power <- predict(model,p)
    print(Power)
  })

}
shinyApp(ui = ui, server = server)
```

Appendix 6: Measured Data for three weeks

Power	Irradiance	modtemp	ambtemp	Humidity
798	693	38.4	21.3	56
999	838	35.1	21.4	52
720	685	35.2	22.8	51
1111	906	36.7	23	50
310	238	31.4	22.8	49
396	315	35.1	24.2	46
540	427	28.7	22.7	50
608	512	38.4	25.1	45
421	331	33.4	24.3	46
455	359	33.5	24.6	45
434	349	37.1	25.3	44
917	799	37.8	21.1	62
761	593	34.4	21.8	59
954	881	49.5	22.8	57
989	913	39.1	22.8	55
1040	1007	35.5	24.7	50
1197	1056	37.5	25.2	44
1085	1066	48.6	26.3	40
1197	1097	39.3	26.3	37
988	942	44.5	27.2	34
962	913	41.8	27.5	32
913	848	40.6	27.5	33
681	557	36	22.8	54
994	931	40.3	24.1	51
591	507	40.3	25.6	45

850	856	41.1	26.5	43
985	958	44.5	27.1	37
1033	1039	46.8	28	36
1064	1061	46.4	29	35
982	1037	42.2	28.7	34
388	317	35	27.5	35
386	301	31.8	27.5	33
269	211	31	27	35
427	336	29.5	21.3	55
489	414	33.6	22.2	51
511	409	33.2	22.7	51
970	916	40.9	23.9	47
240	190	33	23.9	45
649	543	32.8	24.3	44
823	692	43.9	26.7	37
385	314	35.4	26.3	35
291	234	34.1	25.5	35
388	308	31.6	25.7	37
467	360	32.5	25.8	37
263	204	25.1	20	59
323	245	25.7	20.6	57
328	250	26.1	20.8	54
551	439	30.5	21.7	52
347	265	28.3	22.3	48
549	406	28.2	22	50
526	411	30.4	22.6	45
1185	984	33.7	24.5	44

414	345	39.6	24.8	43
543	475	34.3	24.9	43
320	252	34.3	25.7	38
814	694	38.3	20.4	63
789	668	39.5	21.2	59
853	745	39.2	21.6	58
874	797	42.9	23.4	54
900	818	43.1	24.6	50
998	941	42.5	25.9	47
631	588	38.2	26.7	41
766	921	41.9	27.3	33
926	876	43.2	28	33
933	857	40.8	28.5	32
334	289	43.4	29.4	29
759	623	33.8	20.3	65
811	698	39	22.7	59
868	781	42.3	24.5	55
895	813	45.5	25.9	49
859	825	48.8	27.4	44
644	532	42.7	27.5	35
985	1023	46.5	28.1	36
563	557	42.2	28	35
1049	953	39.7	28.9	35
200	161	37.4	29	34
763	671	41	29.4	35
460	352	27.7	20.1	72
467	380	36	21.9	66

293	223	26.9	21.1	67
1017	885	32.2	21.2	68
927	897	44.9	23.3	62
333	264	34.2	26.6	53
875	764	32.7	24.3	56
1084	966	36.6	24.5	55
329	261	33.8	25.3	53
342	266	30.9	24.8	54
214	166	30.5	24.6	54
403	309	26.2	19.2	77
489	393	33.4	21.2	67
540	443	34.7	22.2	63
388	304	31.5	22.5	62
240	185	29.6	22.4	62
577	481	36.4	24.1	60
333	263	34.4	24	58
366	290	38.2	25.1	54
438	385	42.7	26.3	50
307	238	31.5	25.1	53
277	214	29.4	24.6	54
780	656	35.1	20.6	67
828	722	40.2	21.9	64
913	829	41.7	24	57
952	832	40.1	24.4	52
478	411	40.5	25.7	50
537	459	37.4	26.4	48
291	240	36.5	25.6	51

578	482	34.5	25.3	53
147	119	36.4	26.7	46
316	246	26.8	23.6	56
157	122	29	24.1	54
850	720	32.6	21.9	62
620	545	38.9	23.8	55
651	543	38	24.8	51
1011	991	44.9	26.2	48
912	863	43.2	26.8	44
721	647	43.1	27.1	43
962	869	42.5	28.2	42
640	584	39.6	27.9	41
727	642	37.9	28	42
960	963	50.3	30.4	36
437	416	44.1	29.3	36
106	77	16.5	14.6	85
140	104	17.6	15.1	83
188	139	18.5	15.3	82
220	167	20.4	16.5	78
185	141	20.6	16.8	76
487	385	25.1	22.6	58
575	491	34	21	59
322	261	32.5	21	58
318	246	24.7	20.1	63
525	428	30.9	22.1	56
742	644	39.1	23.2	51
237	175	18.8	16.1	88

453	356	27.6	19	78
438	354	31.2	20.6	69
226	173	26.6	19.6	70
423	336	32	21	64
212	166	33.9	22.6	60
517	414	31.4	21.7	62
526	442	38.4	23.5	55
604	504	39	24.5	51
390	325	39.6	25.1	48
379	318	40.3	25.1	48
811	707	32.9	21.4	50
889	802	22.9	22.2	48
998	951	35.7	23.5	47
448	392	41.1	24.3	44
335	304	43.2	25.6	42
637	625	33.6	25.9	40
429	377	42.9	27	36
437	415	43.7	27.8	34
351	299	35.7	27.5	35
957	864	34.7	26.3	37
250	203	30.8	25.7	39
73	52	15.5	14.6	94
111	80	16.5	15	93
205	154	17.7	15.4	90
294	223	21.3	16.2	87
247	185	20.3	16.7	85
306	235	24.8	18.1	81

452	356	27.4	19.8	71
319	247	27.4	19.8	68
160	123	27	20.1	66
313	237	27.5	20.1	65
235	178	26.6	20.6	64
509	376	22.9	16.9	83
523	406	28.8	18.4	79
538	445	37.1	20.5	67
794	692	37.4	22.6	59
369	293	31.9	26.4	50
314	246	32.1	22.7	54
380	293	31.2	22.5	53
474	379	31.3	23.9	52
1132	1118	42.8	25.2	47
903	851	43.1	25.3	44
663	570	44.8	26.2	42
78	56	16.3	15.9	88
102	73	17	16.3	85
124	90	18.1	16.8	84
73	53	18.9	17.2	81
286	211	20.2	17.2	83
426	330	30.4	21.4	70
505	401	30.8	20.5	66
445	348	30.9	21.5	62
327	257	33.4	22.3	58
105	79	26.1	22.1	57
267	201	26.5	21.1	61

418	318	29.9	19.1	68
506	404	32.7	20.2	65
585	491	35	21.2	60
339	262	28.3	20.7	63
342	264	29.7	20.8	63
545	448	34.1	21.8	59
492	422	33.8	21.9	59
249	195	28.6	22.2	57
334	259	28.2	21.9	58
418	327	29.6	22.7	55
204	155	25.4	22	57
239	178	21.7	16.6	76
266	199	22.3	17.2	73
256	192	22.1	17.3	72
347	263	25.6	18.7	68
332	251	25.9	18.5	68
485	382	27.9	20	63
352	266	23.5	20.1	62
285	213	27.6	19.9	63
603	475	30.1	21.2	58
513	429	42.7	23	53
316	287	32.3	22.5	52
124	88	15.7	14.6	98
259	188	17.5	15.2	97
111	79	16	15	95
511	371	20.2	15.4	95
162	117	19.9	15.8	92

296	220	22	16.4	89
335	243	19.6	16.5	86
653	508	26.4	18	79
206	151	21.2	18.1	78
680	520	23.4	18.7	77
1006	905	36.4	20.9	67
194	141	19.6	17.3	76
260	193	22.8	18.6	70
354	266	26.1	19.6	67
646	548	31.8	21.4	59
466	355	33.1	21.3	57
1209	1081	33.6	22.4	53
453	351	30.9	23.3	53
552	517	40.6	26.4	47
516	414	34.6	24.3	49
933	857	40.4	25.3	47
852	765	43.7	25.1	47

Journal paper submitted to the Ethiopian journal of science and technology

Photovoltaic (PV) system performance forecasting and modelling using real -time observation and weather data

Mwende Rita. [1*], Sebastian Waita [1], Okeng'o Geoffrey[2]
[1] Condensed Matter Research Group, Department of Physics University of Nairobi, Kenya
[2] Astrophysics & Space Science Research Group, Department of Physics University of Nairobi, Kenya

ABSTRACT

Increase in demand for energy has led to exploitation of solar energy which is abundant. However, PV power output is greatly influenced by weather parameters. Hence proper modelling and assessment of the impact of environmental parameters on the PV system performance is essential. In this study, a detailed performance analysis of a 1.5 kW PV system installed at the School of Physical Sciences building, University of Nairobi was done to study the effect of solar irradiance, temperature and relative humidity on the power output. A weather station was setup on the site to provide real-time measurements of the ambient temperature and relative humidity. Solar irradiance was measured using a HT304N reference cell and the PV module temperature measured using a PT300N temperature sensor. The current and voltage values of the solar PV system were obtained using a current-voltage solar (I-V) analyzer. Data collection was done daily between 10:00 a.m. to 3:00 p.m. EAT at 30 minutes' interval for a period of 21 days. Support Vector regression (SVR) and random forest regression (RFR) models were built and used to forecast real-time PV power output of a 1.5kW solar PV system. SVR model coupled with principal component analysis outperformed RFR model with root mean square (RMSE) of 45.10, testing dataset compared to RMSE of 94, obtained for RFR model. This shows that SVR model is more robust than RFR and has capabilities of reducing errors during computations.

Keywords: Machine learning; PV performance; weather parameters
DOI:

INTRODUCTION

The availability of sufficient, affordable, and reliable energy is crucial for the wholesome development of any nation. Due to the increasing world populations and advancements in technology that require power, the World's energy consumption is anticipated to rise by 50% by the year 2050 as per the International Energy Agency(Energy, 2019). To date, fossil fuels are the world's main energy source producing about 85% of the world's energy. However, they are non-renewable, unsustainable and have negative environmental impact~~s~~, resulting to a rise in the greenhouse gases, degradation of ecosystems, changes in weather patterns, rising sea level and melting of glaciers (IES, 2019).

[*] Corresponding author: mwenderita74@gmail.com

The focus of renewable energy has shifted to solar energy due to its abundance and availability. Due to her location at the equator, Kenya receives an abundance of solar energy averaging between 5-7 sunshine hours and 4-6 kw/m^2 insolation daily (Solargis, 2017). Photovoltaics are highly popular source of solar energy because they require low maintenance, silent and clean energy (El-Ghonemy, 2012). However, solar power generation is heavily dependent on weather variability. PV power output is influenced by other weather parameters such as temperature, relative humidity, dust accumulation and wind speed (Touati *et al*., 2016);(Khandakar *et al.,* 2019). This inherent fluctuating nature of solar energy poses a major challenge in the quest to fully integrate solar energy power plants into existing power grids without compromising on the stability of the power output. Therefore, as the number of solar PV systems and solar grid connected solar power plants installations increase, there is an urgent need to carry out research aimed at developing techniques and models with the capability of performing accurate real-time site-specific performance analysis and power output forecasting.

The key methods used for PV power forecasting include; physical, statistical and hybrid methods. Physical methods are implemented by predicting the position of the sun and assuming that the other relevant climatic conditions negligible and are used to predict solar irradiance which is in turn utilized to predict PV power output on assumption that solar irradiance is directly proportional to PV power output. Physical PV forecasting methods are broken down into two namely: numerical weather prediction (NWP) and satellite sky imagery. Statistical methods use past irradiance data and PV power output on notion that the future will be a repeat of the past conditions. Statistical methods are divided into two categories; persistence models (classical methods) which assume the future is same as previous value obtained and time series or machine learning models that use historical data over a certain time frame to make predictions i.e. hours, days or even months. Hybrid methods merge two forecasting techniques to enhance predictive accuracy for instance NWP can be used with Artificial Neural networks. Hybrid methods have been adopted to predict power generation data from solar PV modules.

Statistical methods for PV power forecasting have gained popularity because they are data driven. Statistical methods are much simpler to implement, require less input data than traditional methods hence have low computational costs (Malvoni and Giorgi, 2017). They also study the relationships between the weather variables and can determine variable importance (Al-Dahid, 2018). The existing prediction models for solar irradiance use climate parameters like solar irradiance sunshine hours, cloud cover, temperature data retrieved from satellite observations and or ground stations (Salcedo-Sanz *et al.,* 2014). However solar radiation forecasting is mainly used on assumption that solar radiation is directly proportional to PV power output (Wu & Lou, 2011). This assumption results in inaccurate forecasts hence researchers have shifted their focus on directly predicting PV power output using solar irradiance, temperature, wind speed, cloud cover and relative humidity (Khandakar et al., 2019). This approach has yielded forecasts that have a higher accuracy than solar irradiance only based prediction models (Madhiarasan and Deepa, 2017).Khandakar *et al*., (2019) developed a model to monitor, assess and quantify photovoltaic performance using different weather parameters namely wind speed, solar irradiance, ambient temperature, dust accumulation and PV temperature. Prediction models developed utilized Artificial neural network linear regression, M5P decision tree and Gaussian process regression model, they used the data collected on site for the study. ANN prediction model output power was the best. They discovered that incorporating correlation feature selection and relief feature selection together with the ANN yielded more accurate power output predictions.

Isaksson and Conde, (2018)benchmarked on comparing several machine learning techniques with time series models across five distinct locations in Sweden. They concluded that time series model was complex due to non-stationary nature of energy output whereas machine learning techniques were easier to implement. The time series model was also not able to capture different seasonal trends. The ANN and Grade boosting regression performed best compared to KNN and Lasso. It was noted that NWP data was not for the exact location of installations hence reduced accuracy of prediction. They recommended building of different models for different seasons pointing out that using data for a year or more may generalize the model hence yield inaccurate predictions.

Kanwal, (2018) also developed a Gaussian Process Regression and Support vector regression algorithms (SVR) to forecast PV power using temperature and irradiance in Pakistan. The data was trained, tested, and validated. The SVR based trained model was more accurate than GPR. However, the model did not study the effect other weather parameters on PV performance. Abuella and Member, (2016) presented PV power forecasts for 24 hours ahead using twelve variables. The support vector regression, ANN and multiple linear regression models were created. The SVR forecasting model was reported to have performed best compared to multiple linear regression model and at the same time, more robust than ANN.

Kim et al., (2019) developed a PV power forecast model using a selection of unknown weather parameters with confirmed weather forecast data. Support vector machines, K-nearest neighbors, adaptive boosting, random forest regression, classification and regression (CART), and artificial neural network prediction models were developed. The results demonstrated that the random forest regression (RFR) achieved an R-square value of 70.5% which was the best model. However, the model was over-fitted leading to over prediction on certain days. This was attributed to the physical location of the power plant.

Data quality, proper data collection and preprocessing is essential for a good solar PV power predictive model. Isaksson and Conde, (2018) noted that using numerical weather prediction irradiance forecasts data as an input is not ideal since it fails to provide data for the exact coordinates the PV systems were installed. The weather forecast agencies also tend to report after a few hours which can lead to data inconsistency. (Almohri *et al.*, 2014) ;Khandakar *et al.,* (2019); Isaksson and Conde, (2018); (Yerrapragada, 2013)) noted that implementing feature selection technique improved the accuracy of machine learning model by reducing redundancy in the data.

Several concerns about training models with data acquired over large periods of time were raised because they resulted in inaccuracy in the model. Isaksson and Conde, (2018) recommended training models on data according to seasons. They noted that data collected over a year or more can cause inaccurate predictions due to changes experienced in every season. As a result, it is suggested that data obtained during a specific season be used to train a model to predict power output for the same season, which would improve the predictive model's accuracy. Hence exploration of statistical methods based on onsite seasonal data collection and implementation of features selection can be used to build flexible and adaptable short-term PV power forecast methods.

This work aimed at creating interactive models that accurately predict PV power output using real-time observations and weather data using support vector machine and random forest, and analyzed the relationship between weather variables and their influence on the performance of a solar PV system

MATERIALS AND METHODS

Support vector machines

Support vector machines (SVM) is used to perform both classification and regression. It involves the construction of a separation hyperplane or collection of hyperplanes to execute regression on high dimensional data. When the algorithm gets labeled training data it forms the optimum hyperplane which separates new sample data with main the goal being to find a hyperplane $f(x)$ that has most deviation (ε) from the training data and should be as flat as possible (Smola and Olkopf, 2004).

Hyperplane $f(x)$ is expressed by the linear equation

$$f(x) = w_i x_i + b \qquad (1)$$

where b is the slack variable

In SVR, the set absolute error or deviation from the hyperplane should be less or equal to the specified margin called the maximum error ε whose value parameter can be tuned to achieve high accuracy in a model. To ensure the flatness, one has to ensure w is has small as possible this is done by optimizing the problem to give (Smola and Olkopf, 2004).

$$Min\frac{1}{2}|w|^2 \qquad (2)$$

subject to

$$\begin{aligned} &y_{i-}w_i x_i + b \le \varepsilon \qquad (3)\\ &w_i x_i + b - y_i \le \varepsilon \end{aligned}$$

where ε is the deviation, Min is minimize.

Most case errors may occur beyond the ε we denote the deviation from the margin as ξ_i, Equation (3.25) now expressed as shown below (Smola and Olkopf, 2004).

$$Min\frac{1}{2}|w|^2 + C\sum_{i=1}^{l} \quad \xi_i \qquad (4)$$

Constraints

$$\begin{aligned} &y_i - wx_i - b \le \varepsilon + \xi_i \qquad (5)\\ &w_i x_i + b - y_i \le \varepsilon + \xi_i \end{aligned}$$

where $C > 0$ is the penalty parameter of the error term.

When C increases the tolerance for points outside the ε also increases and as C approaches zero the tolerance approaches zero (Smola and Olkopf, 2004).

It has gained popularity because it can effectively classify non-linear data by mapping inputs into high-dimensional feature spaces even when the datasets are small. Kernel function enables one to locate a hyperplane in the higher dimensional space without elevating computational cost. Increase in the dimension of data leads to a rise in the computational cost. When dimension increases and the separating hyperplane is not found in a particular dimension, a kernel is expected to shift the data to a higher dimension support vector classifier. This is achieved by adding a kernel trick which transforms the classes into a higher dimensional space, where classes can be linearly separated (Smola and Olkopf, 2004). Kernels are classified into linear, polynomial, radial basis function kernels. Function (ϕ) maps training vectors (x_i) into higher dimensional space, this is known as the kernel trick $K(x_i, x_j)$ expressed by the equation (Smola and Olkopf, 2004)

$$K(x_i, x_j) \equiv \phi(x_i)^T (x_j). \qquad (6)$$

Furthermore, SVR is less prone to overfitting issue, after training the prediction phase is rapid and work well with high dimensional data these, features make SVR more practical.

Random Forest Regression

Random forest regression (RFR) involves growing of trees depending on random vector Θ_k such that the tree predictor $h(x, \Theta)$ takes a numerical value. A random forest is built by taking an average over k of the trees to reduce the variance hence finding a balance between the two extremes which is expressed as (Breiman, 2001)

$$RFR = \{h(x, \Theta_k), \qquad k = 1, \ldots\} \qquad (7)$$

where $\{\Theta_k\}$ is the random vector and $h(x, \Theta)$ is the tree predictor

Random vectors $\{\Theta_k\}$ are independently identically distributed and each tree selects the most popular class at input x vectors (Breiman, 2001). The mean squared generalization error (GE) for predictor is $h(x)$ is given by the equation

$$GE = E_{xy}(Y - h(x))^2 \qquad (8)$$

where E_{xy} is expected value

The GE for forests converges as to a limit as the number of trees increases. For an accurate RFR model low correlation between residuals and low error trees are key (Breiman, 2001). The more the number of trees the more robust the forest becomes. The RFR do not over fit data as more tress are added but GE is produced.

Experimental setup

This study was on a 1.5kW PV string installed at the Department of Physics, Chiromo Campus, University of Nairobi. It consisted of six 250W Polycrystalline Solinc solar panels connected in series. The solar panels were first cleaned using a clean cloth and plain water before commencing with the measurements. Solar irradiance was measured using a HT304N Reference Cell while the PV module temperature was measured using a PT300N temperature sensor. A current-voltage (IV) analyzer was used to measure the current and voltage of the solar PV system. The data was collected from 10:00 a.m. to 3:00 p.m. EAT at 30 minutes' interval and data analysis was done using R–software and Origin 9.1 software. The data obtained from the PCA was divided into training, testing and validation dataset using the ratio 60%, 20% and 20% respectively.

The training dataset was used to build the Support Vector regression (SVR) with radial basis kernel model in the R software. The penalty parameter C was set at 0.25, 0.5 and 1, the sigma was set at 0.05. Random Forest Regression (RFR) model of 20 trees was built. The validation data was then used to evaluate the performance of the model. Cross validation (CV) was done using the *leave one out cross validation (LOOCV)*, on the training dataset, in order to prevent the models from overfitting. Test set was finally used to determine the performance of the model.

Accuracy Metrics for the evaluation of prediction models

Several metrics are used to determine the accuracy solar (PV) prediction models based on ML techniques. They include mean squared error (MSE), coefficient of determination (R^2), Adjusted R^2 and mean absolute error. The MSE measures an average value of the squares of errors, expressed in the equation (Kim *et al*., 2019)

$$MSE = \frac{1}{B}\sum_{i=1}^{B}\left(y_i - y_p\right)^2 = RMSE^2 \qquad (9)$$

where y_i is the i-th actual value, y_p is the predicted value for y_i, B is the number of samples, and RMSE is the square root of MSE. When the RMSE decreases the predictive model's performance increases.

Mean absolute error (MAE) is the average difference between the predicted and real values, it is computed using the equation

$$MAE = \frac{1}{B}\sum_{B}|y_i - y_p| \qquad (10)$$

The MAE shows measure of errors between the predicted values and the real values but does not indicate the direction of the error.

Coefficient of determination (R^2) is the proportion of the variance of the dependent variable which the independent variables describe as expressed in the equation (Kim *et al*., 2019)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - y_p\right)^2}{\sum_{i=1}^{n}(y_i - y_m)^2} \qquad (11)$$

where y_m is the mean of the actual values of y_i.

The adjusted R-squared is a modification of R-squared and only increases when the independent variable is significant. It is expressed in the equation (Kim *et al*., 2019)

$$Adj\, R^2 = 1 - (1 - R^2)\frac{B - 1}{B - p - 1} \qquad (12)$$

where p is the total number of independent variables.

RESULTS AND DISCUSSION

Data Analysis

Table 1 shows a correlation matrix containing Pearson's coefficient correlation values between power output, solar irradiance, ambient temperature, module temperature and relative humidity.

Table 1: Correlation matrix showing the degree of correlation between the measured weather parameters and power output

	Power output	**Solar Irradiance**	**Module temp**	**Ambient Temp**	**Relative Humidity**
Power output	1	0.99	0.7	0.49	-0.48
Solar Irradiance	0.99	1	0.73	0.53	-0.51
Module Temp	0.7	0.73	1	0.85	-0.79
Ambient Temp	0.49	0.53	0.85	1	-0.95
Relative Humidity	-0.48	-0.51	-0.79	-0.95	1

The correlation coefficients shown in Table 1 greater than 0.70 indicate high correlation between weather variables hence indicate redundant information in the data which often decrease accuracy of a predictive model. The Principal Component Analysis (PCA) allows the extraction of important information from data and presenting them in a set of new uncorrelated features known as principal components. The new data is presented as linear combinations of the original variables with the information in a given PC corresponding to its total variance.

According to the PCA, the directions with the biggest variations are the most significant.

PC1 which is the first PC direction is along the largest variation while PC2 is the second most important with direction orthogonal to PC1 axis. PC1 is a linear combination of variable S_1,$S_{2,.}$ S_k as shown in the equation (Ho, 2019)

$$PC1 = a_{11}S_1 + a_{12}S_2 + \ldots a_{1k}S_k \qquad (13)$$

It can also be written in matrix form as

$$PC1 = a^T S$$

where a^T are the weights vectors

Principal component analysis was used to remove redundant information from the measured weather parameters resulting into four uncorrelated principal components (PC's) as shown. The first (PC1), second (PC2), third (PC3) and fourth principal (PC4) components explained 79.86%, 15.53%, 3.70% and 1.113% of variance of the data respectively. The PCA results indicate that the first two principal components (PC1 and PC2) account for majority (95.19 %) of the variability of the dataset.

3.2: Random Forest Regression

The leave cross validation yielded the best model with highest R^2 of 0.96 and the lowest MAE and RMSE of 51 and 65 on training dataset as shown in Table 2. Figure 1 below shows the measured power output against the predicted power output by RFR model on the training dataset. The regression plot shows high correlation between the measured power output and the predicted power output with correlation coefficient (P) of 0.98. The Adj R^2 of 0.95 shows a good fit of the RFR model on training dataset. The trained model was tested using the testing data set. The testing dataset consisted of 44 observables and 5 variables. Figure 2 shows high correlation between the measured power output and the predicted power output with correlation coefficient (P) of 0.93. The Adj R^2 of 0.87 shows a good fit of the RFR model on testing dataset.

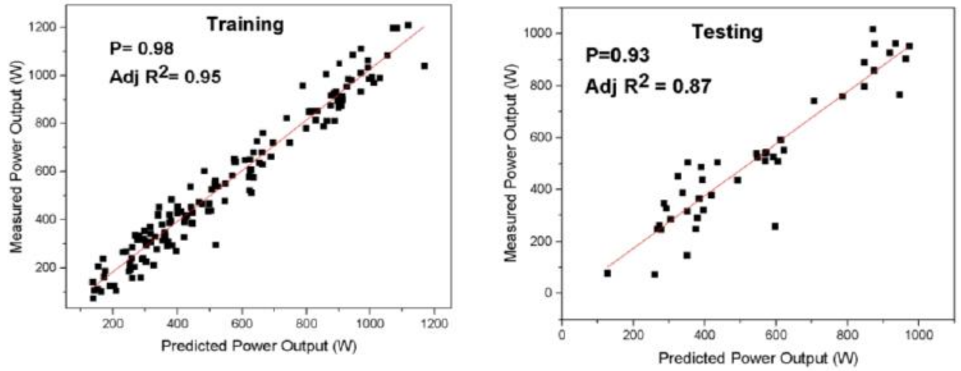


Figure 1:Measured power output against predicted power output by Random forest regression (RFR) model for the training dataset

Support Vector Regression

Support vector regression model based "LOOCV" cross validation yielded the best trained model with R^2 of 0.98 and the lowest MAE and RMSE of 29.01W and 40.4W respectively. Figure 2 below shows the measured power output against the predicted power output by SVR model employing the "LOOCV" technique on the training dataset. The regression plot shows high correlation between the measured power output and the predicted power output with correlation coefficient (P) of 0.99. The Adj R^2 of 0.98 shows a good fit of the SVR model on training dataset.

The regression plot shows high correlation with correlation coefficient (P) of 0.98. The Adj R^2 of 0.97 shows a good fit of the trained SVR model on the test dataset.

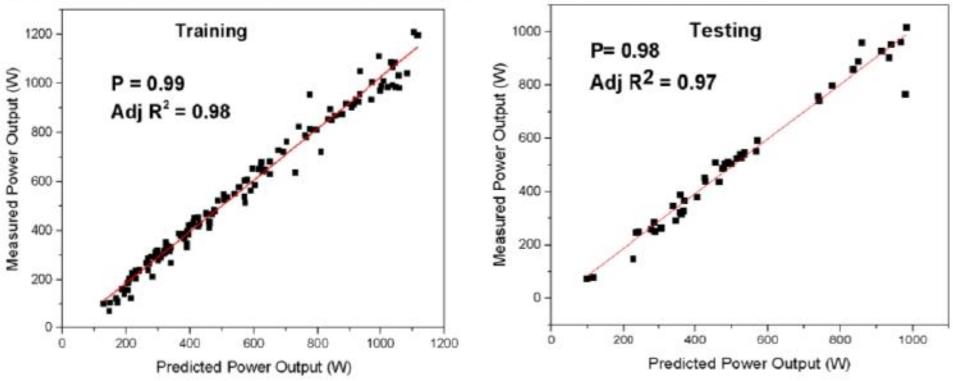


Figure 2:Measured power output against predicted power by Support vector regression (SVR) model for the training dataset.

Table 3 shows that SVR model outperformed RFR model yielding lowest RMSE of 45.1 and highest Adj R^2 of 0.97 on testing dataset.

Table 1:Comparison of performance of RFR model and SVR model on test dataset.

ML technique	**RMSE**		**Adj R^2**		**MAE**	
	Train	**Test**	**Train**	**Test**	**Train**	**Test**
RFR	65	94	0.95	0.87	51.8	68
SVR	40.4	45.1	0.98	0.97	29.01	29.27

CONCLUSION

Real-time performance data of the 1.5kW PV system was collected and studied. The study focused on the influence of solar irradiance, ambient temperature and relative humidity on performance of photovoltaic system. From the results it was observed that solar irradiance had a very strong positive linear relationship with the power output generated by the PV system. However, relative humidity increased with decreasing PV power output showing very weak relationship. PV power forecasting models coupled with PCA were built using SVR and RFR and were successfully trained, validated and tested to forecast real-time PV power output. SVR model employing LOOCV yielded the best model compared to RFR. This is a significant step towards realizing a site-specific and dynamic solar PV performance analysis and forecasting technique.

ACKNOWLEGEMENTS

We acknowledge the support from the Department of Physics, University of Nairobi for providing the equipment used in this research, and the University of Nairobi postgraduate scholarship for the funding towards this research.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

Abuella, M., & Member, S. (2016). *Solar Power Forecasting Using Support Vector Regression.*

Al-Dahid. (2018). *Extreme Learning Machines for Solar Photovoltaic updated the. Figure 1.* https://doi.org/10.3390/en11102725

Almohri, H., Du, C., Hu, Z., & Wang, J. (2014). *Data Analysis and Prediction of Power Generated by Photovoltaic Systems. II*, 22–24.

A.M.K. El-Ghonemy. (2012). Photovoltaic Solar Energy : Review. *International Journal of Scientific & Engineering Research*, *3*(11), 1–43.

Breiman, L. (2001). *Random Forests.* 1–33.

Energy, U. S. (2019). *International Energy Outlook 2019.*

Ho, S. M. (2019). *PRINCIPAL COMPONENTS ANALYSI S (PCA). December.*

IES. (2019). *Environmental Impact of Fossil Fuel.* 4964706.

Isaksson, E., & Conde, M. K. (2018). *Solar Power Forecasting with Machine Learning Techniques.* 46.

Kanwal, S. (2018). Support Vector Machine and Gaussian Process Regression based Modeling for Photovoltaic Power Prediction. *2018 International Conference on Frontiers of Information Technology (FIT)*, 117–122. https://doi.org/10.1109/FIT.2018.00028

Khandakar, A., Chowdhury, M. E. H., Kazi, M.-K., Benhmed, K., Touati, F., Al-hitmi, M., & Gonzales, A. J. S. P. (2019). *Machine Learning Based Photovoltaics (PV) Power Prediction Using Different Environmental Parameters of Qatar.*

Kim, S.-G., Jung, J.-Y., & Sim, M. (2019). A Two-Step Approach to Solar Power Generation

Prediction Based on Weather Data Using Machine Learning. *Sustainability*, *11*(5), 1501. https://doi.org/10.3390/su11051501

Madhiarasan, M., & Deepa, S. N. (2017). *Review of Forecasters Application to Solar Irradiance Forecasting*. *2*(2), 26–30.

Malvoni, M., & Giorgi, D. (2017). ScienceDirect ScienceDirect Forecasting of PV Power Generation using weather input data - preprocessing The 15th of Forecasting PV Generation using weather input techniques Assessing the feasibility of using the Paolo Maria function for a Grazia district . *Energy Procedia*, *126*, 651–658. https://doi.org/10.1016/j.egypro.2017.08.293

Salcedo-Sanz, S., Casanova-Mateo, C., Pastor-Sánchez, A., & Sánchez-Girón, M. (2014). Daily global solar radiation prediction based on a hybrid Coral Reefs Optimization - Extreme Learning Machine approach. *Solar Energy*, *105*, 91–98. https://doi.org/10.1016/j.solener.2014.04.009

Smola, A. J., & Olkopf, B. S. C. H. (2004). *A tutorial on support vector regression ∗ ̈*. 199–222.

Solargis. (2017). *Kenya_PVOUT_mid-size-map_156x220mm-300dpi_v20170921 (1)*.

Touati, F., Al-Hitmi, M. A., Chowdhury, N. A., Hamad, J. A., & San Pedro Gonzales, A. J. R. (2016). Investigation of solar PV performance under Doha weather using a customized measurement and monitoring system. *Renewable Energy*, *89*, 564–577. https://doi.org/10.1016/j.renene.2015.12.046

Wu, C., & Lou, Y. (2011). *Predicting solar generation from weather forecasts*. 528–533.

Yerrapragada, J. B. (2013). *Short-Term Power Forecasting of Solar PV Systems Using Machine Learning Techniques*. *July*, 1–5.