ELECTRIC POWER LOAD FORECASTING USING MACHINE-LEARNING FOR SUBSTATION, TIET, PATIALA

An interim capstone report submitted in partial fulfilment of the requirement for the award of the degree of

Bachelor of Engineering

in

Electronics and Communications Engineering

Submitted by

AAKARSHAN AGARWAL	101706001
NITIN SAPRA	101706111
PARAM PRASHAR	101706112
PRANAV SRIRAM	101706117

Under Supervision of

Dr. Alpana Agarwal

Professor

Department of Electronics and Communication Engineering

Department of Electronics and Communication Engineering

THAPAR INSTITUTE OF ENGINEERING AND TECHNOLOGY, PATIALA, PUNJAB

December, 2020

CERTIFICATE

This is to certify that the project report on, **"Electric Power Load Forecasting using Machine-Learning for Substation, TIET, Patiala"** being submitted by Aakarshan Agarwal, Nitin Sapra, Param Prashar and Pranav Sriram with the help of our mentor Dr. Alpana Agarwal for Capstone Project is a bonafide record of work carried out by us in conformity with the rules and regulations of the institute. The results presented in this report have not been submitted, in part or full, to any other University or Institute for the award of any degree or diploma.

Dated 14/12/2020

DECLARATION

We hereby declare that the design principles and working prototype model of the project entitled Electric Power Load Forecasting using Machine-Learning for Substation, TIET, Patiala is an authentic record of our own work carried out in the Electronics and Communication Department, TIET, Patiala, under the guidance of Dr. Alpana Agarwal during 6^{th}-7^{th} semester (2020).

Roll No.	Name	Signature
101706001	AAKARSHAN AGARWAL	
101706111	NITIN SAPRA	
101706112	PARAM PRASHAR	
101706117	PRANAV SRIRAM	

Counter Signed By:

Faculty Mentor

Dr. Alpana Agarwal

Professor

Electronics and Communication Engineering Department,

TIET, Patiala

Dated: 14/12/2020

ACKNOWLEDGEMENT

We would like to express our special thanks of gratitude to our helpful and kind mentor Dr. Alpana Agarwal who gave us the golden opportunity to do this remarkable project on the topic "**Electric Power Load Forecasting using Machine-Learning for Substation, TIET, Patiala** " which not only helped us gain a deeper understanding of our topic of interest but also helped us research on it and build it.

We would further like to thank Dr. Rajendra Nigam and Dr. Chandan Kumar from Substation and Dr. Pankaj Sinha, from finance dept. for very graciously providing us with the required raw data for our project. We would like to extend our gratitude and appreciation to Dr. Neeru Jindal for guiding us throughout the project and introducing us to effective scientific research and reporting.

Dated 14/12/2020

ABSTRACT

Electric Power Load Forecasting (EPLF) refers to the prediction of electric power consumption in electric power systems. It is an essential process in the planning of future electric power demand and operation of such systems. Accurate and efficient forecasts can lead to substantial savings in operating and maintenance costs, better reliability of power supply and delivery system. This study aims to implement long term forecasts which are usually carried out in a period ranging from a month to a few years, at various sections of TIET, Patiala. Historic electric meter readings will be used to perform training and testing analysis on regression models and other algorithms, which will further assist in predicting future power requirements. Python3 is employed to implement the above model as a user friendly, GUI based software to assist power administrators. The system is protected by facial recognition features implemented through Computer Vision using Raspberry Pi. This would help us to visualize the changing trends of electricity consumption which our campus faces. Proposed work will be compared with existing techniques to prove the efficacy.

TABLE OF CONTENTS

LIST OF FIGURES IN PROJECT REPORT

LIST OF TABLES IN PROJECT REPORT

CHAPTER-1 INTRODUCTION

The main objective of this project is to compare, the performance of different regression techniques on several metrics, run over historic electric load data from our university campus. This would help us to visualize the changing trends of electricity consumption which our campus faces. Further we wish to employ the best-found algorithm in prediction of electric load values, which would be further fed into a GUI based software so that effective predictions can be made and new data can be added to our existing database

1.1 PURPOSE AND SCOPE

Frequent electric power outages in TIET due to the growing population is a major problem for students and faculty. Power supply is often shut down in many parts of the campus, causing inconvenience. Although a 66kV grid has been installed in the area recently, students in hostels face restrictions on AC and geyser timings, seasonally. To plan TIET's power distribution and efficiently use its current electric load capacity, in accordance with growing student populace and in order to minimize power outages by predicting future electric power needs, we propose a project on Electric Power Load Forecasting using machine learning algorithms.

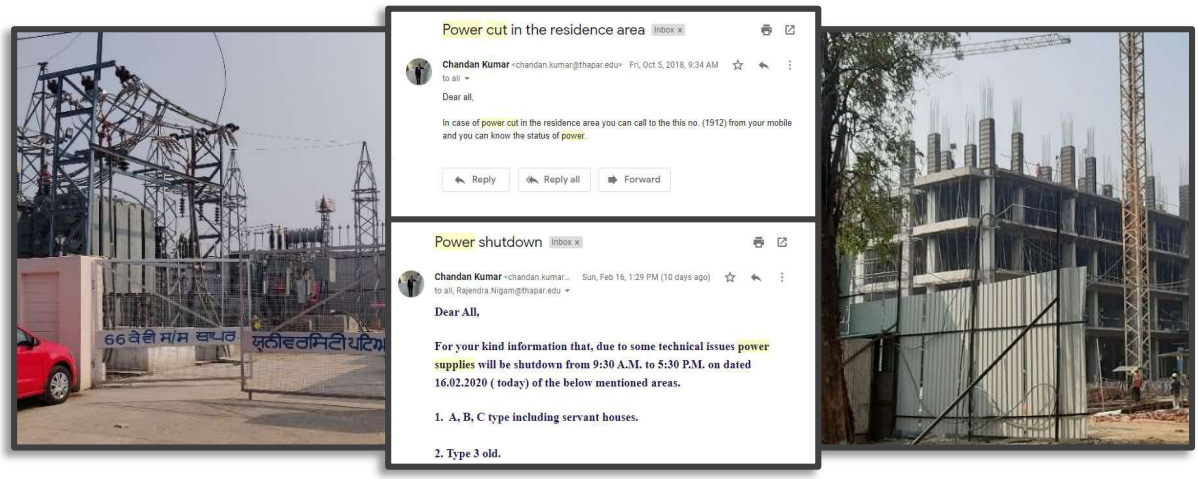


Figure 1: Mails showing electric power situation at TIET.

Due to the fact that electric energy is yielded and utilized concurrently, forecasting the electric load and acquiring economical electric power are imperative for reliable supply. Specifically, a university or college campus, is one of the largest energy utilizing institutions due to the large population that it caters to, and tends to have a broad variation of load which is determined by time and meteorological conditions. For these reasons, a precise electric load prediction method

that can forecast energy utilization in real-time is required for optimum power supply and control Although a wide range of factors impacting power utilization in commercial spaces have been discovered, further studies aimed at increasing the accuracy, efficiency and quantitative prediction of the data. As previously mentioned, accuracy remains a big constraint in providing reliable information in the said case.

Ensemble methods are a domain of machine learning. These techniques incorporate several base models in order to generate one efficient and much more predictive model. Almost every machine learning algorithm aims to reach results which are closest to real life scenarios. Rather than investing our faith in one model and hoping this model brings best results, users can use the technique of ensembling. These methods take a multitude of models into account, and centre those models to produce one terminal model. Using this approach in mind, we propose to design a machine learning model with considerably higher accuracy and lower MSEs than singular models. Through this, we plan to compare and choose relevant models to suit the available dataset and design a user-friendly interface with authentication through face recognition and a login system.

Load forecasting remains a key step in any study aimed at planning and maintenance. Forecasting enables technicians to understand the trends in load behaviour for the present as well as the future. With prior planning and prediction, an organization can save large amounts of money which are wasted and all maintain electricity throughout the day without facing power shut downs. Indian electric meters use kWH i.e. kilo Watt Hour as the basic unit.

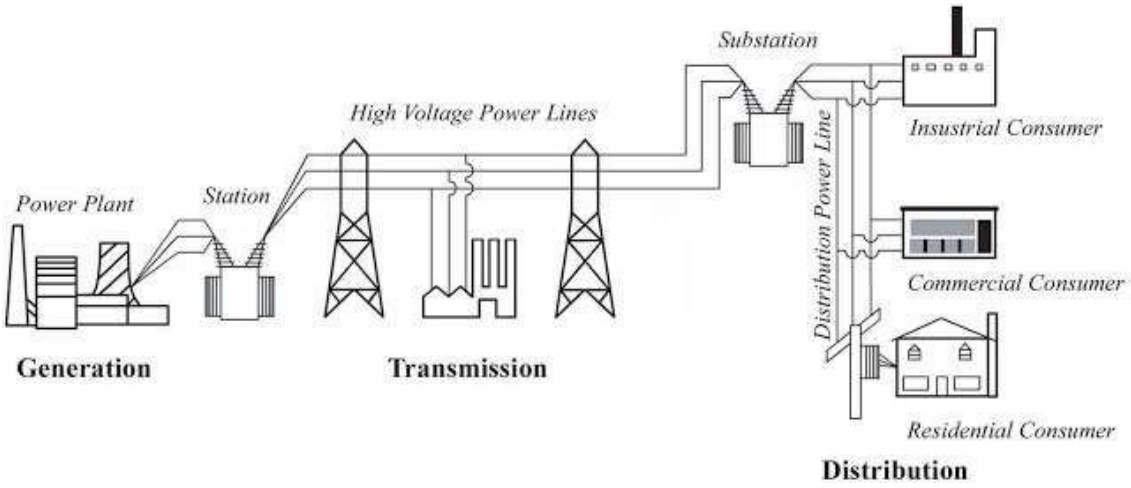


Fig 2: Supply of electric power to consumers (http://www2.econ.iastate.edu/)

The above diagram is a schematic representation of the steps involved in receiving electric supply at personal or commercial spaces. While personal spaces like houses and also small businesses can sustain without any hassle. Large commercial places like an IT park, university campus, etc. often use grids for the sole purpose of receiving uninterested power supply. Due to this it is quite beneficial to plan and schedule cycles of generation of energy so that energy is available when it is needed.

1.2 APPROVED OBJECTIVES

I. Obtain and add relevant variables to pre-process electric load data.
II. Building robust machine learning models to forecast electric load values.
III. Comparing regression models based on relevant key performance indicators.
IV. Build an attractive user interface for the user to comfortable operate and calculate values.
V. Securing the software with face recognition system.
VI. Deploying the software on RaspberryPi for hassle-free usage.

CHAPTER 2 LITERATURE SURVEY

2.1 THEORY ASSOCIATED WITH PROBLEM AREA

Load forecasting is categorized into 3 types. These are namely short term, medium term and long-term forecasts. Each of these types has a different application and is summarized in the table below:

Type	Lead time	Aim
STLF	½ hour-few hours	Maintenance and planning
MTLF	Days-weeks	Organize and plan Seasonal peaks
LTLF	Months-Years	Planning generation growth

Table 1: Existing load forecasting categorization

In this project report, long term forecasting is used as basis for prediction. Long term predictions by nature always have the disadvantage of being inaccurate. This is due to the fact that LTLF predictions are always dependent upon temperature effects. Sometimes historic meteorological and economic data are simply unavailable for a long-term duration, making it difficult to analyse and make an effective dataset. Also, based on predictions, it is a huge task and takes a big investment to upgrade existing facilities. Even if we have surplus energy at a time, storing it using present technology is a cumbersome task.

2.2 EXISTING SYSTEM AND SOLUTIONS

Extrapolation:

This approach involves fitting a trend on a said curve for making prediction outside the given dataset. Curve fitting function is chosen by load variation, obtained from the pattern of prvious data. Some important curve fitting functions include:

- Straight line: $w = a + b*x$
- Parabolic: $v = a + b*x + cx^2$
- Exponential: $z = c\ e^{dx}$
- Double exponential: $z = \ln^{-1} (a + c\ e^{dx})$

Using the method of least square final predictions are made. Usually, extrapolation is not very dependable, even if good results are obtained, due to lack of confidence in the technique. In order to make concrete decisions based upon this method, a constant and reliable dataset has to be used.

ARIMA:

ARIMA is an acronym for Autoregressive Integrated Moving Average. ARIMA in a single variable, is a forecasting technique that predicts the succeeding values of a series based entirely on its own weight. Its major use is in the area of STLF, needing at least 40 historical data points. It works best when dataset exhibits a stable or consistent pattern over time with a low number of outliers. Also known as Box-Jenkins (after the original authors), this technique is better than exponential smoothing when the data is larger and the correlation between previous

observations is stable. If the data is small or highly unstable, then some smoothing method may perform better. In case the dataset is smaller than 38 data points, some alternate method must be considered

ARIMA algorithm begins with checking Stationarity, which implies that the series remains at a fairly consistent pace with time. If inconsistency in data plotting exists, as in most financial or commercial projects, then your dataset is said to be non-stationary data. The data should also reflect a stable variance with time. This is often observed with a dataset that is highly seasonal and heightening at a rapid rate. In such a situation, the troughs and crests in the seasonality will become more amplified over time. If stationarity condition in not satisfied due to some reason, many functions used in this analysis simply fail to work.

Differencing:

If non stationarity is seen during plotting of the taken data, then differences in values are to be observed. This is an effective way of converting a nonstationary series to a stable one. This is corrected by differencing the observation in the current interval from the consequent historic one. If such conversion is performed just once in the data, the data is called as singly differenced data. This process effectively removes the trend if the taken data is increasing at a consistent rate. If it is increasing at an rapid rate, one can use the same method and subtract the series again. The resultant data is finally said to be second differenced data.

Autoregressive Models:

ARIMA technique aims in detailing the fluctuations in a stationary series as a function of selected parameters. These parameters are called as AR, autoregressive parameters and MA, moving average parameters. A single parameterized ARMA model is represented as:

$X(t) = A(1) * X(t-1) + E(t)$

where X(t) represents the time series being used in analysis

A (1) = order 1 parameter relating to autoregressive nature

X(t-1) = Lag 1 intervals in time-series

E(t) = Error contained by the mode

This implies that any value in X(t) is explained by some function carrying a historic value, X(t-1), added with some incomprehensible random error, E(t). If the predicted value of A (1) was .70, then the actual value of the series would be related to 70% of its value, in the previous interval. Of course, the series could be related to more than just one past value.

$X(t) = A(1) * X(t-1) + A(2) * X(t-2) + E(t)$

The above equation shows that the current value of the series is a sum of the two consequent preceding values, X(t-1) and X(t-2), added with some random error E(t). This becomes an autoregressive model carrying order two.

Neural Networks:

An artificial neural network or ANN defines a set of algorithms designed with an inspiration from the working of biological neural networks and human brain. The use of these algorithms is aimed at recognizing patterns in a given data set. As biological neurons interpret sensory or visual data, ANNs are capable of interpreting such data involved in perception, clustering as well as labelling in the form of numeric, contained in arrays and relating to real world scenarios. Datasets containing images, sounds, textual findings and numerical time series etc are often fed as an input to this network.

Neural networks have a major application in clustering, classification and regression problems. The type of problem can be thought of as a layer placed on top of the data set, which helps the program to identify it and take necessary steps. Unlabelled data is grouped together based on the similarity in data points and thus classes/ labels are formed for further analysis. (Neural systems can likewise segregate attributes that are fed to different models for grouping and ordering; Along these lines, one can consider profound neural systems as parts of bigger AI applications including calculations for reinforcement learning, arrangement and regression.)

Correlation between two distinctly separate and unrelated variables can be calculated using this algorithm. This is often known as a static prediction. Along similar lines, provided with enough data, ANNs can fairly estimate future values, required for prediction with a high amount of accuracy. Regression relating to the past and the future can be run. A label can be given to the data found for future. Deep learning algorithms don’t necessarily involve time, or the fact that something is yet to happen.

Nodes act as the building blocks of any layer. A node, relating to a human neuron, is a unit where calculations and adjustments take place. A neuron in contrast, fires when it encounters sufficient stimuli. A node merges input from the dataset with a set of attributes, or weights, that either intensifies or diminish that input, while allotting significance to inputs with regard to the work, the algorithm is trying to understand. The product of inputs and weights are summed and then this number is passed through an activation function. This is done in order to determine if and to what extent that signal should pass further through the ANN to affect the final result, say, an act of regression/clustering. The neuron is "activated" when signals pass through the activation function.

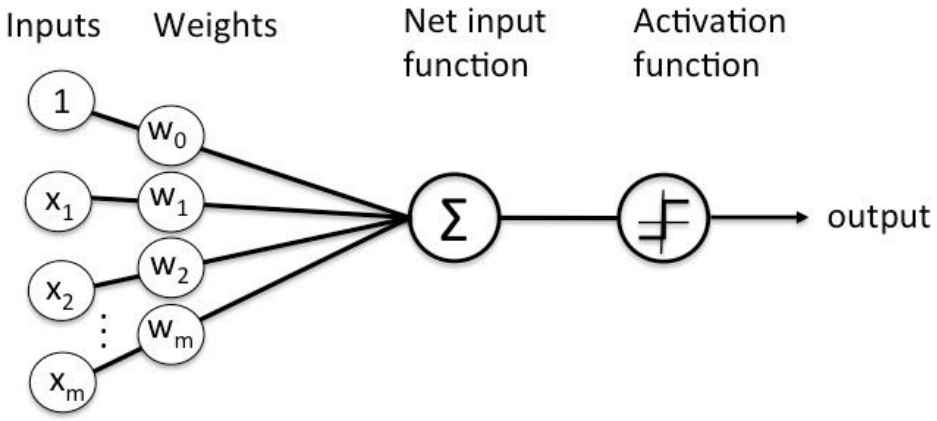


Fig 3: Layout of a neural network

A node layer consists of rows of neuron-like switches which turn on or off according to the input provided, to the network. Output of every layer acts as an input to following layer, and this starts from the layer where inputs are provided by the user.

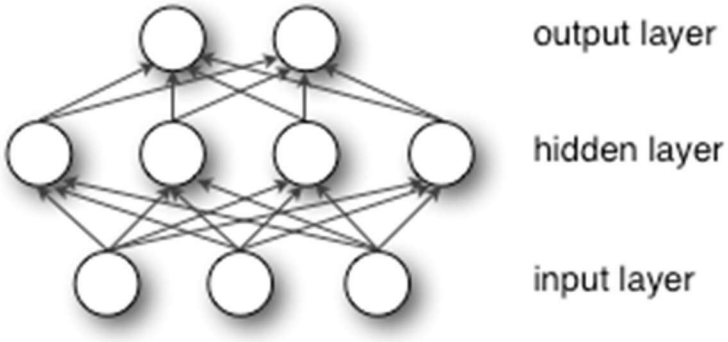


Fig 4: Layers in a neural network

Choosing which attribute contains more significance than others, weights are given by the user which differentiates and chooses how a neural network will work.

Fuzzy logic

The sheer fact that fuzzy logic is very flexible and works rather simply in situations where linguistics and numerical data is processed, it is quite natural to think of fuzzy logic as a means to use this time series analysis. The work on fuzzy time series was started by Song and Chisson in the year 1993 and then modified by Chen in the year1996. The logic behind the usage of this algorithm in time series relates to is to segregation of the universe of discourse from time series in gaps or periods (the fuzzy sets), and understand how each area acts in different conditions. The rules of these models tell how the said periods link with themselves with passing time, as values change from one set to another. In other words: a linguistic variable is often formed. to represent the numeral time series, and these areas will be the linguistic terms of given variable. When a linguistic variable is constructed to represent the universe of discourse, a "vocabulary", is designed, and then the fuzzified dataset is made using, words in that said vocabulary. Then an algorithm learns this set of vocabulary thus formed, using which further calculations are made.

Technique	**Advantage**	**Disadvantage**
ARIMA	Ease of understanding and implementation	Long times for calculation
Multiple regression	Is able to include several parameters with respect to load data and brings more perspective into dataset.	Sometimes it becomes difficult to map functional dependence among variables.

Fuzzy logic	It works better on unresolved and varying real life data. It can combine numeric and linguistic characters for better implementation.	High complexity. Needs specialized teams for implementation.
Artificial neural networks	Combines time series and regression approaches. It is capable of non-linear models' adaptations and does not require any assumption of working relationship between load and meteorological variables.	It is unable to provide insights on the problems fed into the system. It doesn't provide any scope of optimum network selection by the user.

Table 2: Advantages and Disadvantages of existing solutions

2.3 STUDY OF LITERATURE REVIEWS

1) Title: Load Forecasting Techniques and Methodologies: A Review

Author(s): Arunesh Kumar Singh, Ibraheem, S. Khatoon, Md. Muazzam , D. K. Chaturvedi

Year of Publication: 2012

This scientific research paper begins with introducing readers to a supplier's perspective into electrical forecasting. These people have to remain involved in timely production, transferral, dispatch as well as maintenance of markets. So, it becomes a vital task to generate predictions/forecasts well before time. Factors including festivities, weather, meteorology, etc make a huge impact on the consumption and other trends of use in the market. Maintaining and increasing accuracy of current systems and techniques. remains the central idea of this research paper. This paper also introduces the reader to the different types of forecasting and what are the applications of each said type. Further the authors aim to cover three main groups of prediction systems used today. These include traditional, modified traditional as well as soft computing/ machine learning techniques.

The paper starts the algorithmic discussion from simpler methods like regression, least squares, approximations and interpolation/extrapolation. The second category covers stochastic techniques like ARIMA/ARMA as well as SARIMA. Finally, the third type consisting of neural networks, SVMs, trees etc are discussed. The paper also presents pros and cons of each of the discussed technique. The paper concludes by taking into effect, how soft computing/ML methods are gaining more popularity in recent times, owing to the ease of operation, reliability as well as accuracy.

2) Title: Long term electric peak load forecasting of Kutahya using different approaches

Author(s): Y. Aslan, S Yavasca, Celal Yaşar

Year of Publication: 2011

Just like the previous research paper, this scientific research paper takes us ahead on the ways electric system arrangements are done and how these systems are controlled and managed. As this paper was specifically based on long term planning, it gave us a lot of information regarding the same, which otherwise remains less explored.

This paper also talks about the economic aspects of the said planning and focuses on the secure functioning through reliable predictions. Giving us a global perspective, this study is based on the research done in city of Kutahya in Turkey. Authors explore a wide range of algorithms from least squares, regressions to advanced techniques in soft computing like ANNs. Eight years of data from 2000-2008 is considered and attributes such as load, population increase as well as temperature are taken into consideration.

3) Title: Load Forecasting for a Campus University Using Ensemble Methods Based on Regression Trees

Author(s): María del Carmen Ruiz-Abellón, Antonio Gabaldón, Antonio Guillamón

Year of Publication: 2018

This research paper gives the reader, an insight to using ensemble training in load prediction use-case. Based upon the readings taken from a college in Cartagena in Spain, the authors use boosting, random forest, conditional forests and bagging schemes over the data. Usual indicators like temperatures and schedules are taken into consideration while performing this STLF study.

The study is aimed at helping concerned departments in the college, involved in purchase of electricity by giving them an accurate prediction at least 2 days before the schedule. The paper shows how this study was able to save up to 11 percent of costs involved in the commercial buying of electricity and also presents a model relating the above techniques to Spain's real time electric prices for optimal savings.

4) Title: Predictive artificial neural network models to forecast the seasonal hourly electricity consumption for a University Campus

Author(s): JihuiYuanab, CraigFarnham, ChikakoAzuma, Kazuo Emura

Year of Publication: 2018

Exploring further on soft computing techniques for prediction mechanism, we found this paper as a reliable source in understanding usage of ANNs used to carry out a similar research like ours, at a college in Japan. The authors take a wide array of attributes in the dataset to make it

moderately complex, which is quite helpful in reaching better results. Attributes such as irradiance, hourly proportion, dry bulb temperature conditions, hour, week, etc are taken into account. The paper aims to find daily power consumption among the 3 zones of the collef=ge campus.

Methods such as FFNN, LM back-spread techniques, etc are used. The paper concludes the study by comparing R-square values obtained in the process along with other common regression metrics. The authors give a further proposal to stress upon indoor human activity and class schedules to improve this presented model.

5) Title: Electric Load Forecasting with Deep Machine Learning

Author(s): Ala Adin Baha Eldin Mustafa Abdelaziz, Ka Fei Thang,Jacqueline Lukose

Year of Publication: 2019

This research paper begins by presenting the idea that due to growing population and rapid pace of technological advancements, consumption of electric energy has gone up, in recent years. Electricity always remains a vital energy source connecting all pleasures of 21st century to it and it is therefore necessary to develop systems needed in prediction and management of this indispensable treasure. The accuracy of such developed systems should be fairly high, considering the importance of this resource and how essential it is, in even the most basic human activities.

Authors aim to implement different kinds of models in order to achieve highly accurate results and attain an understanding as to what pros and cons does each model possess. The models chosen by the authors include eXogenous input-NARX, feed forward neural network-FFNN, bidirectional long-short term memory-BLSTM, LSTM and MLR.

The selected models are then carefully implemented using MATLAB through its deep learning and neural network toolbox. In addition to training of these datasets, the authors go a step further in developing an android based mobile application. This mobile application allows the user to control, understand and envision the project findings. The criteria used by the authors to compare performances of the taken algorithms is through MAPE evaluation.

The authors have presented a dataset, taking values from all of the world, covering 3 continents i.e. Tasmania, Australia and Canada. The dataset used in the paper covers reading from a whole decade i.e. from 2006-2016.

2.4 COST ANALYSIS

Apart from operational costs the major costs incurred during this project were for the purchase of hardware components required for the face recognition system integrated within the EPLF system.

These components are namely –

Rasberry Pi 3b+ (Rs.3,210)

Camera Module (Rs.379)

Connector Cables (Rs.520)

Peripherals (Rs. 2,270)

For the customer, from simulations and from past research the recorded electricity demand data taken from a substation of power development authority of certain campuses across India from 1st January 2014 to 31st December 2015, it was found that there was an overall power saving of around 3.5 million kilowatt hours during this period depending on average size, population and type of infrastructure on campus.

This power saving can translate into cost cuts in millions of rupees according to the regional per unit rate of power. Furthermore, the cost savings increase not linearly but parabolically year on year as the relative saving increase faster as the years go on.

2.5 RISK ANALYSIS

There are various risks associated with development, testing and execution of this project. They are:

- Electrical failure due to overwork of Raspberry Pi leading to overheating of ICs and eventual permanent damage to the circuit.
- Since all the equipment, especially the camera is very delicate there can be damage during handling or installation of the hardware components.

- Environmental factors such lighting, visibility, obstruction due to obstacles can lead to varying results.
- Human errors in hardware and software can also pose a threat to the project.
- Proper data entry of weather and other corresponding variables.

CHAPTER 3

FLOWCHART

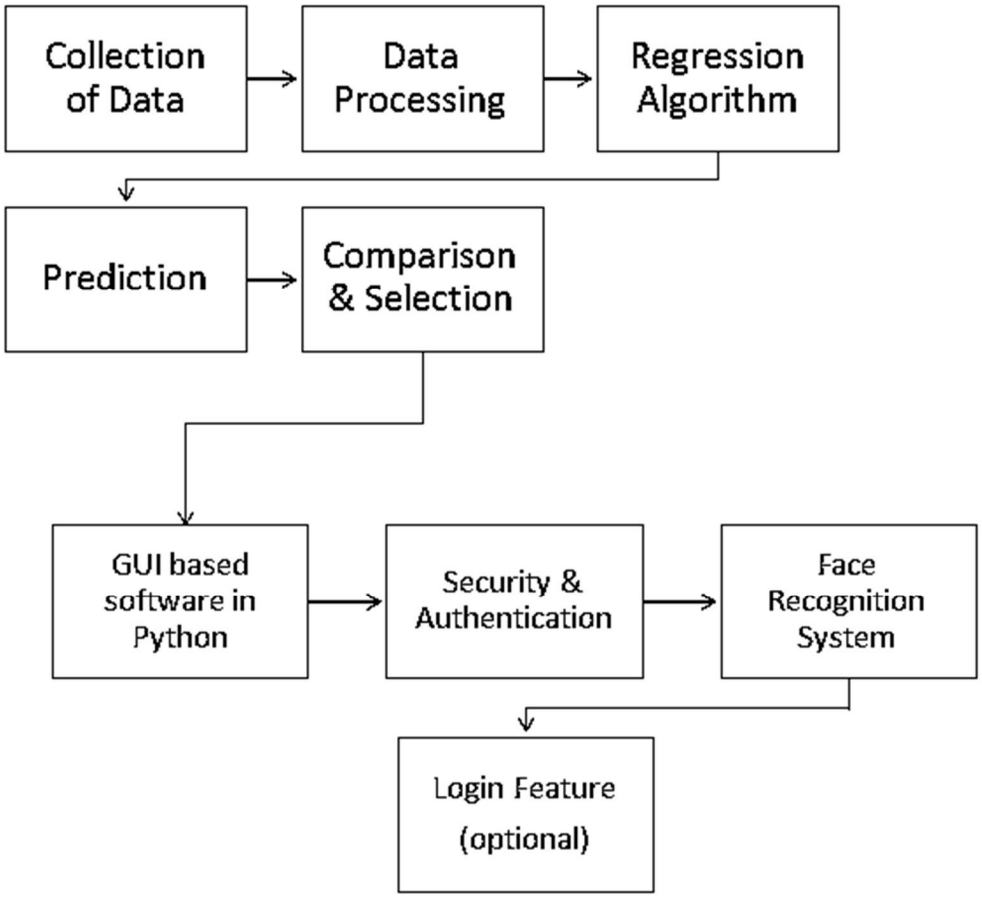


Fig 5: Flowchart of the project process, divided into phases I and II

Process flow:

Phase I

1. The phase I of project flow deals with data procurement and preprocessing. Relevant features are extracted from raw data to prepare a dataset.
2. Other attributes like month, year, average weather, student population and special days are added to enhance the complexity of the dataset.
3. This is followed by setting up Python3 environment.
4. Relevant packages namely, scikit-learn, pandas, matplotlib, numpy, etc are installed and the system is ready to perform regression calculations.
5. Five regression algorithms (explained ahead) are performed and regression metrics are calculated. These metrics are quite useful in comparing the performance of algorithms.
6. This is followed by plotting Actual vs Predicted plots and bar graphs comparing metrics.
7. Best performing algorithm is chosen by the end of this phase.

Phase II

1. This phase retains the best performing algorithm from the previous phase. This would be used as an output to a GUI software, using which a technician can find out predicted values of a particular month of her choice.
2. Preparation of GUI based software using tkinter in python.
3. Connecting the software to firebase for login authentication.
4. Alternatively, providing face recognition as a means for logging in.
5. Implementation of the above processes on a raspberry pi, so that technicians don't face the hassle of running the software on incompatible devices.

Technologies used: Python, Google Firebase, Raspberry Pi, Raspbian, Atom,

Python packages used:

1. Tkinter
2. Datetime
3. Scikit-learn
4. Matplotlib
5. Numpy
6. PIL
7. Pyrebase
8. OpenCV
9. Tkinter.ttk
10. Face_recogntion

CHAPTER 4

PROJECT DESIGN AND DESCRIPTION

PHASE -I

4.1 DATASET AND DATA PRE-PROCESSING

Raw historic data from campus substation was received. Relevant features were carefully extracted from more than 200 excel sheets and compiled together. High and low weather data was added to our dataset, along with corresponding year and months. Another attribute 'special day' was introduced, to segregate exam/vacation months from normal months, as considerable difference in electric load is observed during these months. The database was prepared from Aug, 2010 to Mar, 2019 with 105 entries. The amount of entries remains on the lower side owed to the unavailability of previous entries with the concerned department. However, to increase the scalability and improve the existing performance of algorithms, a function will be added to the proposed GUI, to add further entries.

	A	B	C	D	E	F	G	H	I
1	Snum	Year	Month	Load	Whigh	Wlow	Students	Spl Day	Average
2	1	2010	Aug	549300	34	25		0	30
3	2	2010	Sep	549300	31	22		1	27
4	3	2010	Oct	391481	33	20		0	27
5	4	2010	Nov	353306	30	16		0	23
6	5	2010	Dec	317362	22	9		1	16
7	6	2011	Jan	417224	20	8		0	14
8	7	2011	Feb	339425	15	11		0	13

Fig 6: Visualizing the dataset

4.2 ALGORITHMS USED IN ANALYSIS

4.2.1 Linear Regression

Regression is a method of estimating the target value based on independent forecasts. This method is widely used to predict and determine the relationship between variables. Regression strategies vary depending on the number of independent variables and the type of relationship between independent and dependent variables. In the simplest case i.e Simple linear regression, the model has one predictor variable and there is a linear relationship between the predictor variable (x) and forecast variable(y).

$$y=a_0+a_1x$$

In the case of multiple regressions, the linear relationship between two or more predictor(independent) variables and the forecast(dependent) variable is modelled. It essentially helps to analyse the relationship among multiple predictor and forecast variables.

$$y=a_0+a_1x+a_2x_2+\ldots+a_nx_n$$

4.2.2 K Nearest Neighbour

The KNN comes under the umbrella of instance based learning algorithms. It is often called a lazy learning algorithm owed to the fact that the involved function is only approximated locally and further calculations are done until final function results are received. Normalising the data fed as input to this algorithm considerably increase its accuracy. For regression problems, a technique can be developed to allocate weights to contribution from neighbours. Due to the fact that nearer neighbours impart more to the mean than the distant lying ones. Many a times the reciprocal of the distance of neighbours is assigned as a weight, which makes weighing scheme compatible to the previous point, leading to higher accuracy.

4.2.3 Support Vector Machines

SVMs can be used in analysis for both, regression as well as classification problems. After providing the SVM model set with dataset for each category, this algorithm is able to categorize the new data and put it into respective labelled classes. The support vector machine captures these data points and pulls out a hyperplane (which is just two lines, in case of two dimensions) that separates the label. This line acts as the limit of decisions: anything that falls on its side is categorized as label 1, and anything that falls on the other as label 2.

4.2.4 Gradient Boost

Gradient boosting is an AI method used in analysis of both regression and classification problems, which creates an expectation model as a troupe of frail forecast models, normally decision trees. It constructs the model in a phase wise manner like other boosting techniques do, and it sums them up by permitting streamlining of a discretionary differentiable loss function.

Boosting calculations are depicted as iterative functional gradient drop calculations. That is, calculations that upgrade a cost work over capacity space by iteratively picking a capacity that focuses in the negative inclination bearing. This practical angle perspective on boosting has

prompted the advancement of boosting calculations in numerous zones of AI and insights past relapse and grouping.

4.2.5 Random Forest

The name random forest suggests that this model is made of a forest, which is a collection of decision trees. A normal forest model simply uses mean of all underlying trees, but random forests have certain parameters relating to how results have to be calculated. Building trees in random forest technique requires random sampling of training data points. In addition to this, random subsets of attributes are chosen, while nodes are split during the process of training and finally, random sampling is done over the training observations.

During the process of training, every tree learns from a random sample of data points in the iteration. Drawing of random samples from the larger pool, called bootstrapping is done in the next step. It is always a possibility that same datapoint is used several times during a single iterative run. In the event that we train each tree on an alternate arrangement of tests, we can acquire a lower change in the general resulting forest, regardless of whether the individual trees may require high fluctuation concerning a particular arrangement in the preparation of information.

Towards the end of the process, bagging, short for bootstrap aggregating is done. In bagging, each individual learner is trained on different bootstrapped subsets of the data and then the predictions are averaged.

4.3 METRICS USED IN ANALYSIS

4.3.1 Explained Variance Score

As the name suggests, explained variance score metric is involved in calculation of variance regression score. It measures the proportion in which a model is dispersed on a given data. It is estimated as follows:

$$explained\_variance(y, \hat{y}) = 1 - \frac{Var\{y - \hat{y}\}}{Var\{y\}}$$

where, $\hat{y}$ = estimated target output

and y = correct target output

The optimal score that is desired is 1.0 and lower values are not preferred.

4.3.2 Maximum Error

The maximum error function calculates the difference among two parameters conveying the same phenomenon in worst case scenario. For an absolute fitted regression line, this parameter computes to zero, but this is never the case with real life training data. This metric in fact, provides the extent to which error could occur during analysis. The maximum error is defined as follows:

$$Max\ Error(y, \hat{y}) = max(|y_i - \hat{y}_i|)$$

where, $\hat{y}_i$ = the predicted value of the ith sample

and y_i = the correct value of the ith sample

It is preferred that an error value near 0 is obtained.

4.3.3 Mean absolute error

The mean absolute error function calculates a risk metric relating difference in paired observations, conveying the same phenomenon. Mean absolute error (MAE) estimated over $n_{samples}$ is defined as

$$MAE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} |y_i - \hat{y}_i|$$

where, $\hat{y}_i$ = the predicted value of the ith sample

and y_i = the correct value of the ith sample

It is preferred that an error value near 0 is obtained.

4.3.4 Mean Squared Error

The mean squared error, also known as mean square deviation function calculates a risk metric relating to the expected value of the squared value of loss. The value of this metric remains strictly positive.

$$MSE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (y_i - \hat{y}_i)^2$$

where, $\hat{y}_i$ = the predicted value of the i[th] sample

and y_i = the correct value of the i[th] sample

It is preferred that an error value near 0 is obtained.

4.3.5 Mean Square Log Error

Mean square log error is a logarithmic alteration to mean squared error. It computes the metric relating to the expected value of squared logarithmic losses. It is associated with relative percentage differences between true and predicted values. The lower the value of MSLE, the better is the metric performance of the algorithm. Zero can be the best possible value for the same. Mean squared logarithmic error (MSLE) estimated over $n_{samples}$ is defined as

$$MSLE(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} (log_e (1 + y_i) - log_e (1 + \hat{y}_i))^2$$

where, $\hat{y}_i$ = the predicted value of the i[th] sample

and y_i = the correct value of the i[th] sample

4.3.6 Median Absolute Error

Median Absolute error tells user about how spread out the range of data is. It is a robust measure, which is often used if the data is not in normal form. The error is calculated by taking the median of absolute differences of all values, between the target and the prediction. The lower the value

of MedAE, the better is the metric performance of the algorithm. Zero can be the best possible value for the same.

$$MedAE(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|).$$

where, $\hat{y}_i$ = the predicted value of the i^{th} sample

and y_i = the correct value of the i^{th} sample

4.3.7 R^2 Score

R^2 Score, also called coefficient of determination, tells the user about the proximity of data to the fitted regression line. This is basically an indicator of how well a curve is fit. It is equal to the ratio of explained variation to total variation. The value of this metric lies between 0 and 1, with the higher value as the better performing one. The values of this parameter may go into negative, due to the fact that model could be erratically worse. Zero remains the score for models which provide a constant output, independent to any given input.

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \underline{y})^2}$$

$$\underline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

PHASE -II

4.4 USER EXPERIENCE

As discussed earlier, phase II of the project deals with the creation of a graphic tool protected with login system and face recognition. In this section, we discuss the various details about the development of the same.

4.4.1 Face Recognition

Face recognition system of our project has been created in Python3.7 using libraries OpenCV and face_recognition. OpenCV integrates in itself, modules related to the implementation of "Computer Vision". Computer vision is a branch of AI that helps machines interpret and recognize the visual world. Using this technique, images or videos captured through digital

cameras and mobiles are fed into deep learning models which can precisely identify, classify and recognize objects and patterns. The use of CV extends from face recognition tools to other areas like surveillance, counting people/objects, intelligent vehicle control, anomaly in commercial production, drone control and medical image analysis.

In the particular case applied to our project, the user first enters an image to the database. This given image is used for face matching during further login attempts.

During login attempt image is captured from webcam of the system/ RaspberryPi and face is detected from it. This face is then transformed and cropped out of the whole image. Finally, the cropped image is then fed into a deep neural network to perform recognition.

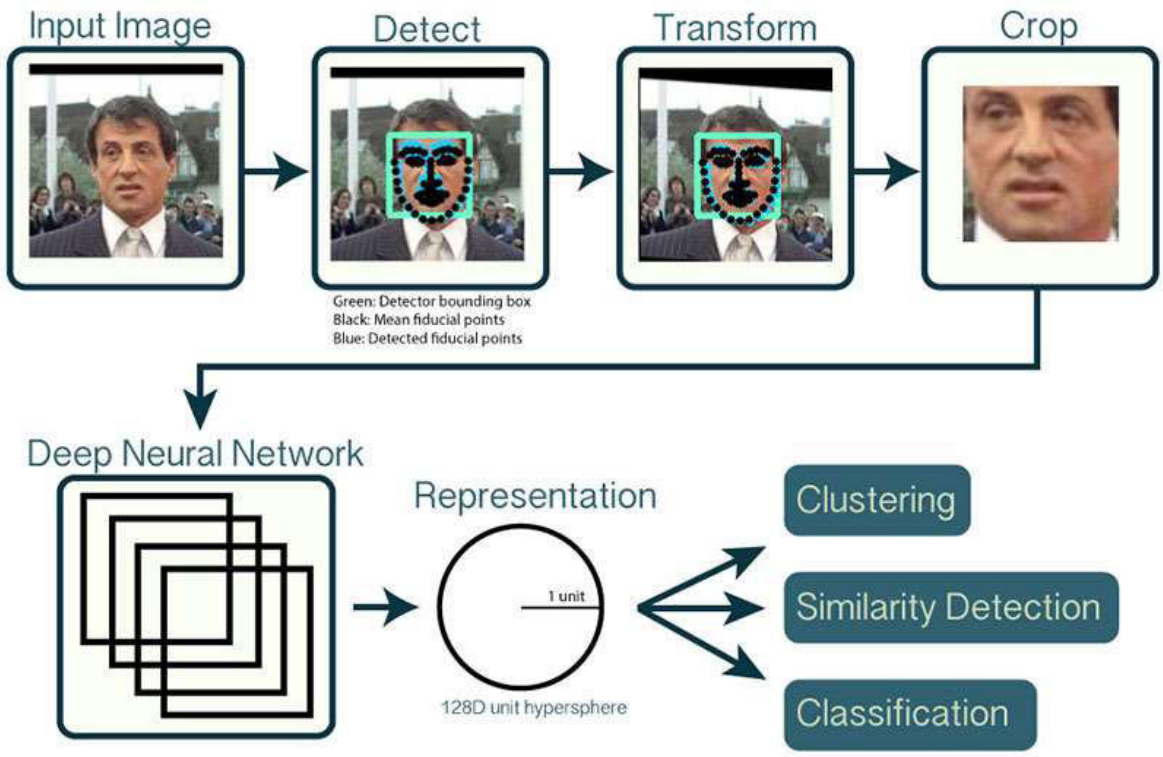


Figure 7: Working of OpenCV face recognition system.

4.4.2 Login window

In case of poor lighting or unavailability of a user, if someone else wants to log in to the system, an alternate is provided . Login window is created in using the module tkinter, which is the default tool for creating user interfaces in Python and comes pre-bundled. Tkinter is used along with its attribute ttk, as tkinter alone is unable to process multiple threads and gives and exception in such a case. An alternate which can be used in such a situation is mtkinter, but was

not brought into use for this particular development. A picture of load forecasting logo is added in “gif” format using the library PIL. PIL can be installed using module “Pillow”, and also comes pre-bundled with the Python distribution.

Two labels and entries are used for displaying the texts and entering of Username and Password of the user, attempting to log into the system. At last button is used to login to further windows. Custom background of “gray25” ois applied to the main windows.

If username and password of the user are authenticated, the user is directed to calculation page, if not, he is shown a message regarding the wrong details entered.

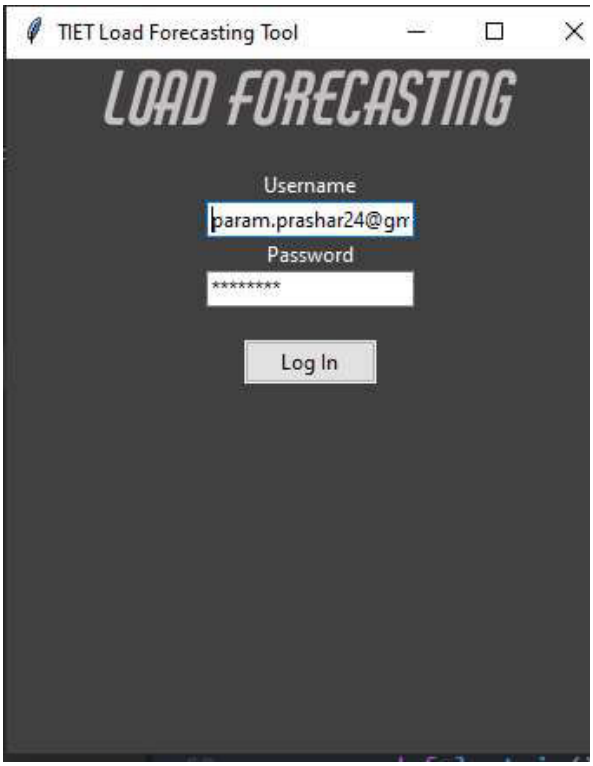


Figure 8: Login GUI

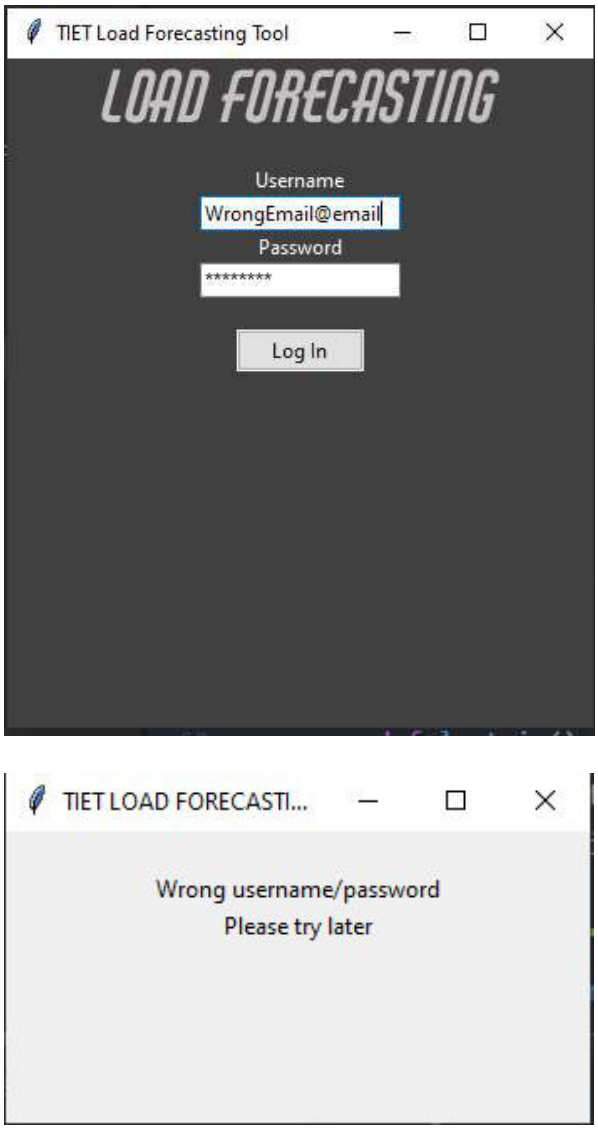


Figure 9: Window after entering wrong details.

4.4.3 Firebase

Firebase is an online service, provided by Google which is often used to create mobile, web and other applications. One salient feature provided by this service is the creation and usage of online databases. The project uses authentication feature of firebase real-time databases. In the first go, user is asked to add email id and password for creation of entry. After the details are provided, an email is generated and sent to the same email id. User authenticates his id through this mail. Now the user is ready to use the same details with the load forecasting software. When user enters his details on the first window explained above, his details get verified and is able to move to further windows.

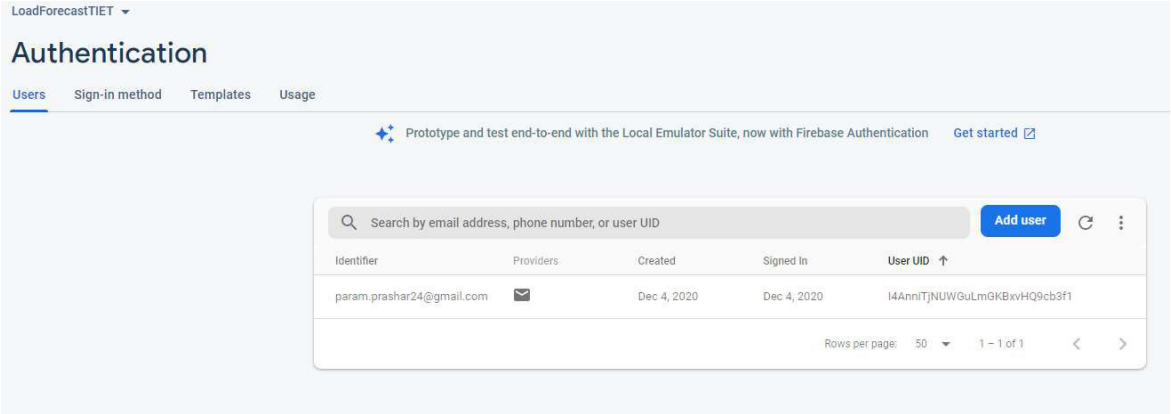


Figure 10: Email entry after authentication at Firebase database

4.4.4 Calculation

After a user successfully authenticates his identity through login system or face recognition system, he is shown the calculation window. Calculation window gives user the ability to choose which forecasting algorithm to calculate through. An options menu or more commonly known drop down list is used for this purpose and a choice between Gradient boosting and random forest can be availed. Two radio buttons are present for choosing if the month is a normal or holiday/examination month. After these two text fields are added where the user can enter Average monthly temperature and the month number. These details are extracted from entry fields using get() function after the Calculate button is pressed. The extracted details are added to an array and using predict() function of scikit learn, the machine is able to calculate corresponding result, which is later shown to the user.

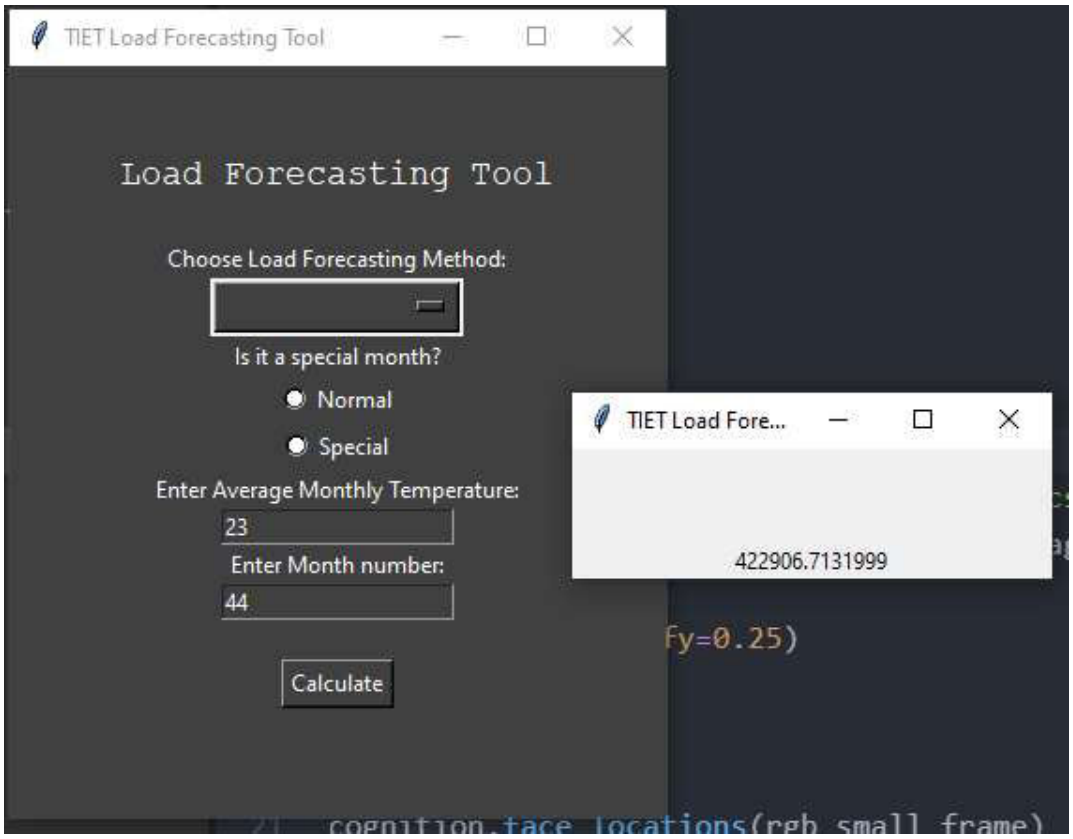


Figure 11: Forecasting tool in action

4.4.5 Raspberry Pi

The last stage of the project deals with deployment of the software over RaspberryPi. This stage was chosen due to the many advantages this embedded microcontroller presents over a regular system. A regular system used at the substation, which is mostly used for keeping Excel databases and/or email purposes is often inadequate or incompatible to be used for a machine learning implementation. RaspbianOS, which is based on Debian provides many advantages over a traditional windows machine for this particular use case. The addition of modules to the system is often hassle free in linux.

Setting up the RaspberryPi, Raspbian OS 2020 is installed on it using a microSD card as the storage. An additional camera hardware is used to properly extract the face recognition capabilities of our software. It is powered by a 15.3W USB powered power supply. The hardware used comes with 2GB RAM and a 64-bit quad core Cortex-A72 processor.

Initial interfacing was done through ssh using PowerShell on windows. The graphical interface of Raspbian can be explored using VNC Viewer, after the IP address of the RaspberryPi is known.

Figure 12: RaspberryPi and camera connected to power

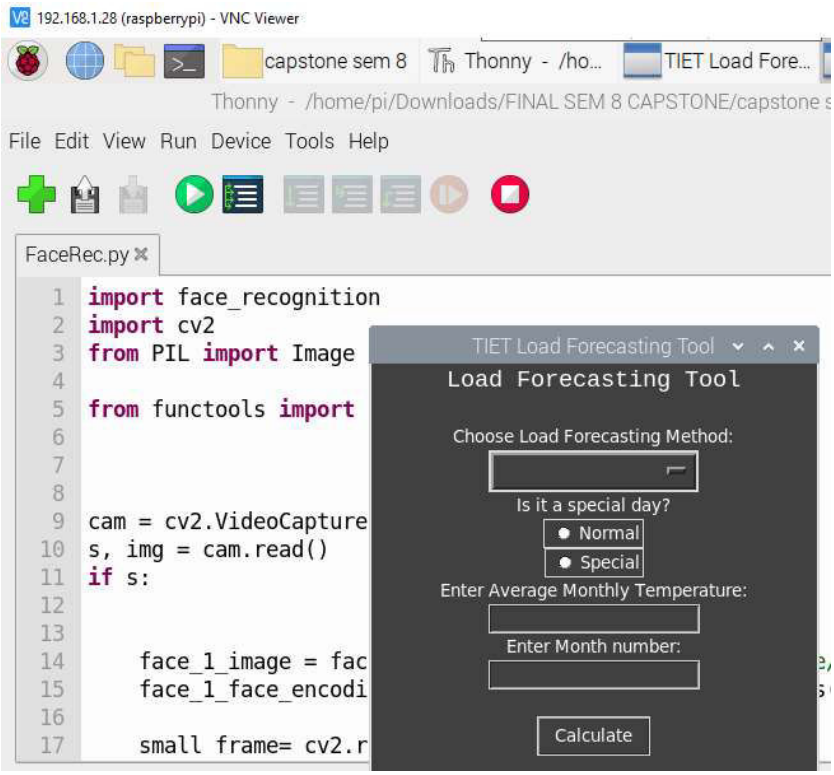


Figure 13: Project deployed on RaspberryPi's RaspbianOS

4.5 RESULTS AND OBSERVATIONS

Algorithm	Explained Variance Score	Max Error	**MAE**	**MSE**	Mean Square Log Error	Median Absolute Error	R^2 Score
Linear Regression	0.69111	0.23435	0.08665	0.01069	0.00022	0.06740	0.69016
Random Forest	0.84727	0.15541	0.06245	0.00530	0.00011	0.05941	0.84613
Gradient Boosting	0.85734	0.17886	0.05668	0.00498	0.00010	0.03922	0.85558
SVM	0.72662	0.19041	0.08447	0.00945	0.00019	0.07234	0.72618
KNN	0.81345	0.22135	0.06512	0.00646	0.00013	0.04659	0.81275

Table 3: Regression Metrics results

Result:

Analysing the results presented above, Random Forest and Gradient Boosting techniques, which are respective subsets of ensemble learning, computed the best possible results. The phase II of this project takes prediction values from the same for further analysis and calculations. This assures the idea presented towards the beginning of this report, about ensemble methods being better performing than others..

4.6 OUTPUT AND COMPARISON PLOTS

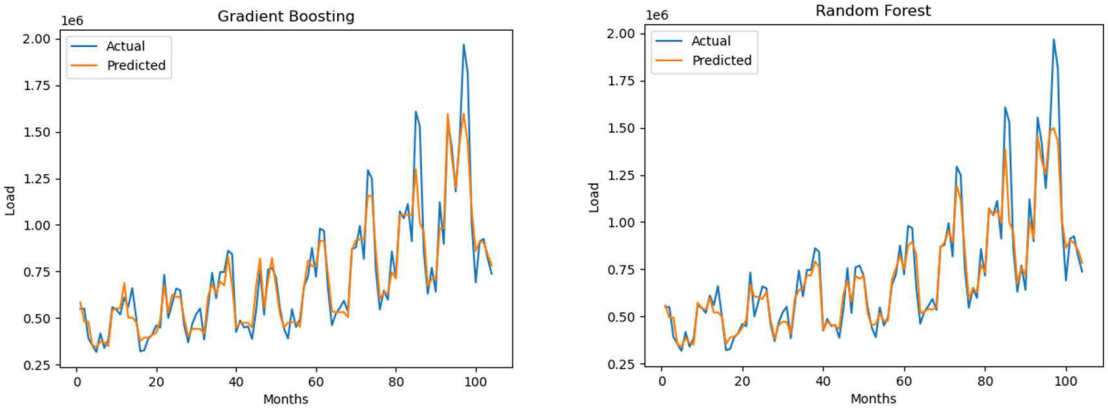
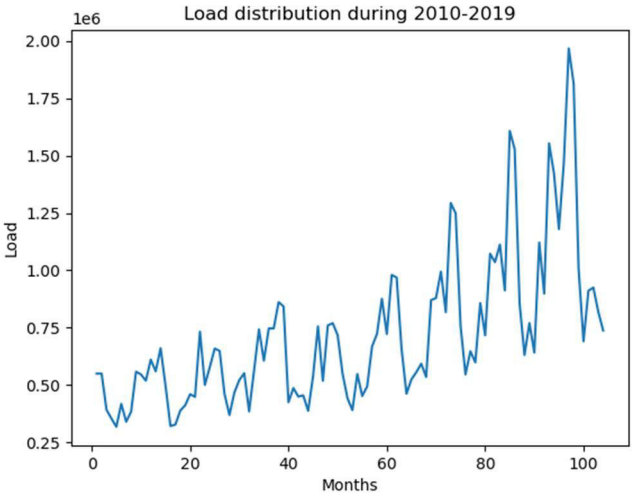
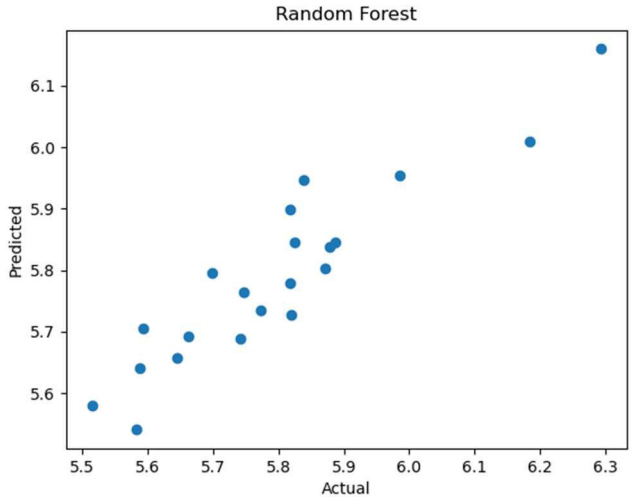
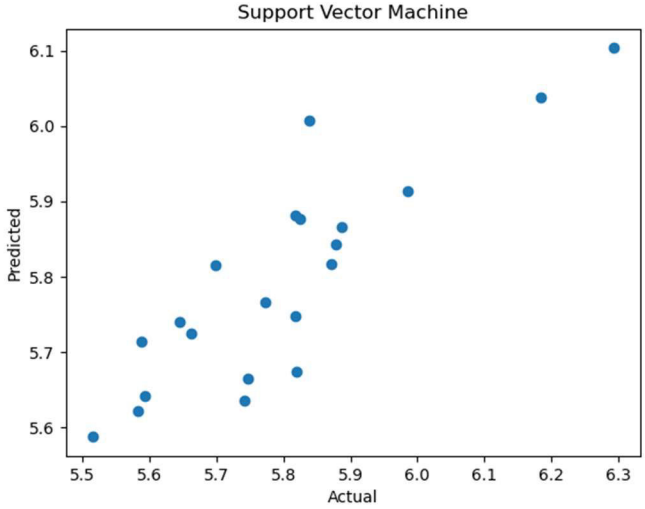
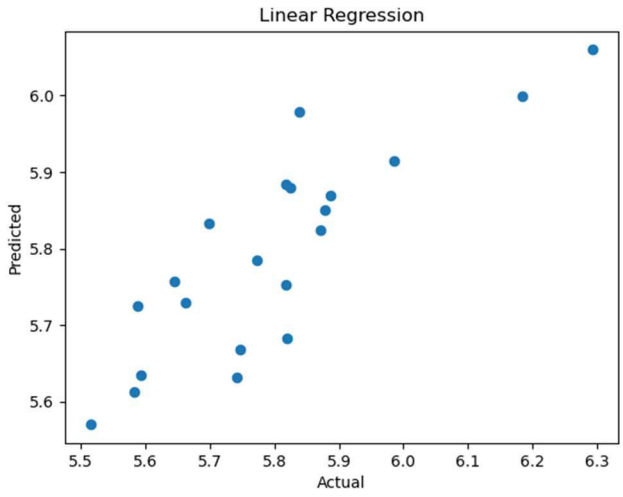


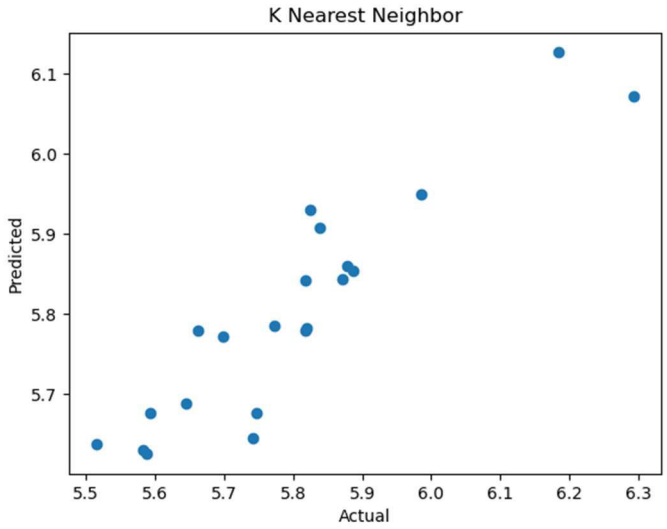
Figure 14: Actual vs Predicted Gradient Boosting and Random Forest fittings.











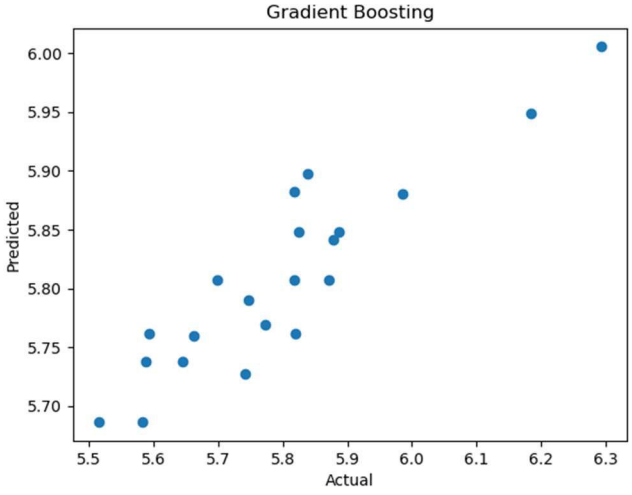


Figure 15: Scattering plots of predicted vs actual values of different algorithms

Figure 16: Regression metrics comparison

Scatter plots often present rich data visualization insights. Predicted vs Actual plots here, present actual values in comparison to those predicted by the given regression algorithms. In such plots, the model attaining values closest to a straight diagonal line is often considered best, in terms of performance. All the points in such a scatter plot lie closer to the diagonal line in

case R2 value of the corresponding plot is large (closer to unity) and vice versa. Looking at the plot of linear regression model, the data points seem to have a random relationship and do not essentially seem to follow a given diagonal line. This can also be verified from the R2 of the same i.e. 0.69016, and the worst performing among the five models taken under consideration.

The first plot presented in Fig 15 is a load distribution. The graph presents the time range of database with monthly load values. The plot shows regular spike of increased consumption during summer months, each year, owing to the high temperature, often reaching up to 40° C. The most obvious factor leading to the huge spikes coming towards the end of the three is the higher yearly intake of students in the campus, leading to construction of new hostels and academic buildings, which in turn leads to a higher consumption in electric energy. Further we can also see, how close the ensemble prediction models are fit, after training over a 75:25 train to test ratio, cross validated dataset.

Bar graphs in Fig 17 compare regression metrics (Explained Variance score, MSE and R2 score) of all 5 algorithms presented in the project, where the better performance of Gradient boosting and Random Forest is clearly observed.

4.7 STANDARDS USED IN THE PROJECT

- 1232-1995 IEEE Standard for Artificial Intelligence and Expert System Tie to Automatic Test Equipment (AI-ESTATE): Overview and Architecture
- 610.4-1990 - IEEE Standard Glossary of Image Processing and Pattern Recognition Terminology
- 208-1995 - IEEE Standard on Video Techniques: Measurement of Resolution of Camera Systems

4.8 UG COURSES USED

- Machine Learning (UEC711)
- Soft Computing (UEC704)
- Object Oriented Programming (UTA018)
- Computer Programming-I (UTA017)
- Electrical Engineering (UEE001)

CHAPTER 5

OUTCOMES AND PROSPECTIVE LEARNINGS

5.1 OUTCOMES

A major outcome of the project was developing a machine learning model based on the concept of regression algorithms which will help in making quick and accurate predictions regarding electrical power load in the future. Adding to this we have an application in which we will have a portal for users to login using two modes: face recognition or login username and passcode. This will help in maintaining security of the system so that the software can only be used by the authorized personnel. Further, by continuously updating the database of load from the substation, the accuracy of model increases. After finishing our project, we would like to conclude the major outcomes of this project as follows:

- Learning the concepts machine learning and regression algorithms
- Learning new methods of facial recognition (through OpenCV).
- Analyzing and testing the working of ensemble model.
- Being able to set up guidelines for the most efficient use facial recognition.
- Understanding concepts of regression metrics.

5.2 PROSPECTIVE LEARNING

The study conducted in the field gives us the significant knowledge to pursue the project. There are some prerequisites if someone wants to work upon on this project. One should be having full information regarding python or MATLAB, how the results are computed and how the error and other metrics are calculated. One should also know how to code in python and install OpenCV to create a local desktop environment. There may be various libraries which are included in the code which help in simplified and easy execution of the algorithm. So, one should have basic information about all the libraries used in the code.

5.3 ENVIRONMENTAL, ECONOMIC AND SOCIAL BENEFITS

Environmental benefits:

By forecasting the exact demand, maximum utilization of power generating plants is achieved. The forecasting avoids under generation or over generation and reduces wastage of power leading to less impact on the environment.

Economic benefits:

The use of electrical load forecasting by power generation companies and plants is highly beneficial for their financial and economic health. Forecasting minimize the risks for the utility company. Understanding the future long-term load helps the company to plan and make economically viable decisions in regard to future generation and transmission investments. Helps to determine the required resources such as fuels required to operate the generating plants as well as other resources that are needed to ensure uninterrupted and yet economical generation and distribution of the power to the consumers. This is important for short, medium, and long-term planning. The load forecasting helps in planning the future in terms of the size, location and type of the future generating plant. By determining areas or regions with high or growing demand, the utilities will most likely generate the power near the load. This minimizes the transmission and distribution infrastructures as well as the associated losses.

Social Benefits:

Everyone everywhere has most likely experienced power outages and the inconveniences that come with it. With the help of electrical load forecasting, not only can we minimize these inconveniences, we can also effectively manage distribution so that areas where power is really needed can get it. For example, it is very important that street lights always get power during the night otherwise if there is a cut even for a short time, it could lead to grave accidents and even loss of life. Furthermore, if we save power in residential and commercial areas, some of the power can be diverted to the slums for helping the people living there. Thus, all of society can benefit from a better management and distribution system.

5.4 REFLECTIONS

During the course of this project, from ideation to completion, the entire team has worked consistently towards the final goal. This project was started with a clear end point in mind however the path to it was murky. There were many roadblocks along the way, especially in the beginning when the challenges seemed colossal and it was easy to question and doubt the importance of this project. Several changes and modifications were made in the overall scheme as work progressed and certain realizations were made. Many hefty ideas that initially seemed easy and viable became painstakingly infeasible. Many long hours were spent reflecting as a team and debating among each other on what to focus on and what to compromise on, However, the final goal remained ambitious and it was delivered to full satisfaction and no compromises were made on the novelty and quality of the project.

5.5 FUTURE WORK PLAN

I. Extending the project to include data from Dera Basi Campus

II. Betterment of Machine Learning Model for future predictions (BUY/SELL)

III. More interactive GUI Integration

IV. Large Scale Deployment

5.6 CHALLENGES FACED

Several challenges were faced during this project from start to finish and bringing it to completion wasn't free of problems.

Many administrational problems were faced in order to get the required data sets and electric load information from the substation. There were other technical problems as well, such as not obtaining the desired level of accuracy form the different machine learning models that were implemented in the initial runs. Other software issues during integration of face recognition system were also faced.

Apart from software, working with the hardware aspect of the project also posed many challenges. Acquiring all components with the desired specifications was a difficulty and was

accompanied by monetary constraints. The installation and wiring of hardware components were also not an easy task.

Many times, problems occurred in the form of ideology differences, human errors, miscommunication and other unpredictable factors. However, the biggest challenge of all, which was not in anyone's control was the occurrence of the covid-19 pandemic and the subsequent nation-wide lockdown. This led to a complete halt on all work towards the project for a while as everyone needed tie to adjust to the new way of life. However, all team members connected over online mode and co-ordinated with each other to overcome even the biggest one of these challenges.

5.7 OUTCOMES ACHIEVED

A. An ability to design a model, system, or process to meet desired needs within realistic constraints such as economic, environmental, social, political, ethical, health and safety, manufacturability, and sustainability.	
A1. Analyze program needs to produce problem definition for electronics and communication systems.	**Yes/No** **Yes**
A2. Carries out design process to satisfy project requirement for electronics and computer system.	**Yes**
A3. Can work within realistic constraints in realizing systems.	**Yes**
A4. Can build prototypes that meet design specifications.	**Yes**

B An ability to function on multidisciplinary teams.

B1. Shares responsibility and information schedule with others in the team.	**Yes**
B2. Participates in the development and selection of ideas.	**Yes**
C. An ability to communicate effectively.	
C1. Produce a variety of documents such as laboratory or project reports using appropriate formats and grammar with discipline-specific conventions including citations.	**Yes**
C2. Deliver well organized, logical oral presentation, including good explanations when questioned.	**Yes**

D. The broad education necessary to understand the impact of engineering solutions in a global, economic, environmental, and societal context.	
D1. Aware of societal and global changes that engineering innovations may cause.	**Yes**
D2. Examines economics tradeoffs in engineering systems.	**Yes**
D3. Evaluates engineering solutions that consider environmental factors.	**Yes**
E. A recognition of the need for, and an ability to engage in life-long learning.	
E1. Able to use resources to learn new devices and systems, not taught in class.	**Yes**
E2. Ability to list sources for continuing education opportunities.	**Yes**
E3. Recognizes the need to accept personal responsibility for learning and of the importance of life-long learning.	**Yes**
F. An ability to use the techniques, skills, and modern engineering tools necessary for	

engineering practice.	
F1. Able to operate engineering equipment**.**	**Yes**
F2. Able to program engineering devices.	**Yes**
F3. Able to use electronic devices and systems modeling software for engineering applications.	**Yes**
F4. Able to analyze engineering problems using software tools.	**Yes**

5.8 LEARNING OUTCOMES FOR CAPSTONE PROJECTS

Course Learning Outcomes	**Rate b/w 1 to 5 (5-Achieved, 1-Not Achieved)**
Developing new/multidisciplinary technical skills	**4**
Using professional and technical terminology appropriately.	**4**
Effectively utilizing and troubleshooting a tool for development of a technical solution.	**5**
Analyzing or visualizing data to create information.	**5**
Creating a technical report with the usage of international standards.	**4**
Acquiring and evaluating information	**5**

MAPPING TO CLO'S OF PROJECT

The purpose behind the project is to give the students a fascinating and fruitful approach to learn multidisciplinary abilities like procuring and investigating information, actualizing

information in the real world and figuring out how to make research reports in an appropriate way. The idea we decided to pursue for our capstone project, satisfies all the centre ideas of the CLO. Bringing information on various subjects into a solitary pool to make a software framework, has surely added weight and experience to our specialization. We got a chance to go through research papers, discover information and insights, break down and assess the scientific information and adequately discover specialized arrangements, which really made us learn a lot from this experience.

MAPPING TO STUDENT OUTCOMES

The project, as planned, has created in us the ability to work in a time constrained condition. We are understanding the methods of scientific research, assessing information and discovering specialized solutions. We would figure out We also hope to learn to give professional e-presentations and posters presentations.

CHAPTER 6

PROJECT TIMELINE

6.1 GANTT CHARTS

The project is still advancing, and we have broadly classified our work into phases. Currently we present the work and research performed from January to May. The overall progress of the project, as well as individual contribution is present in the following Gantt Charts:

PHASE-I

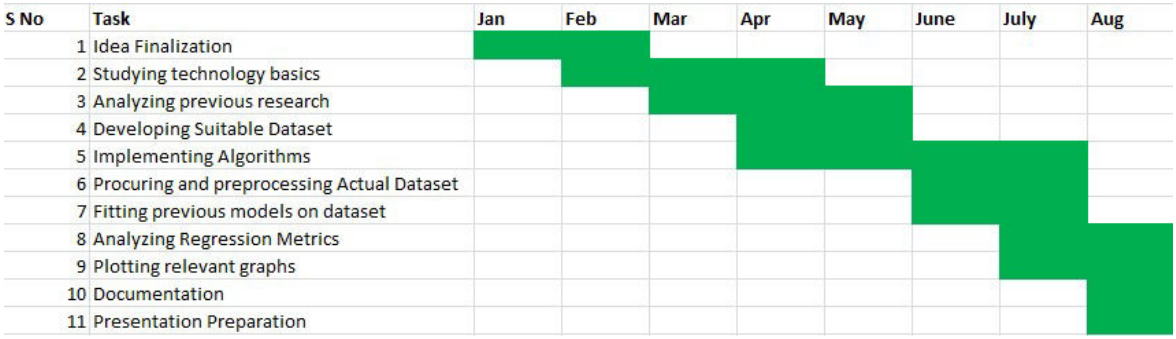


Figure 17: Gantt chart showing team's project timeline

Figure:18 Gantt chart showing Aakarshan's project timeline

:

Figure:19 Gantt chart showing Nitin's project timeline

:

Figure:20 Gantt chart showing Param's project timeline

:

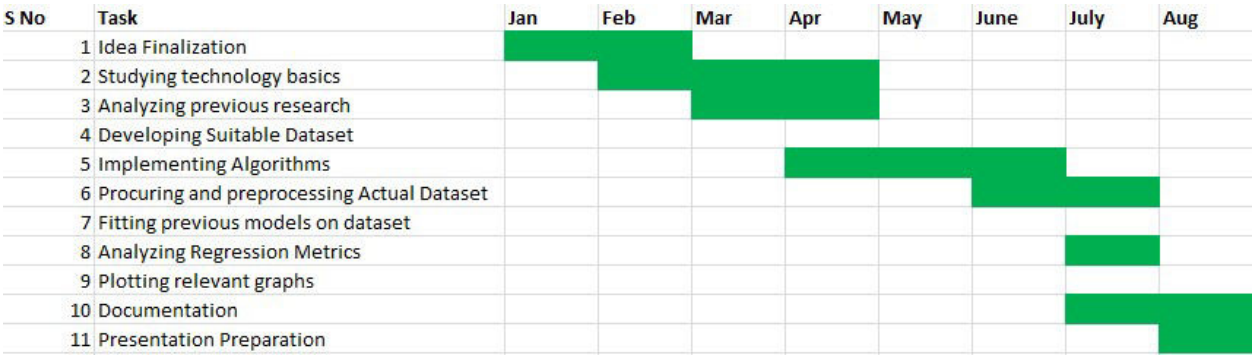


Figure:21 Gantt chart showing Pranav's project timeline

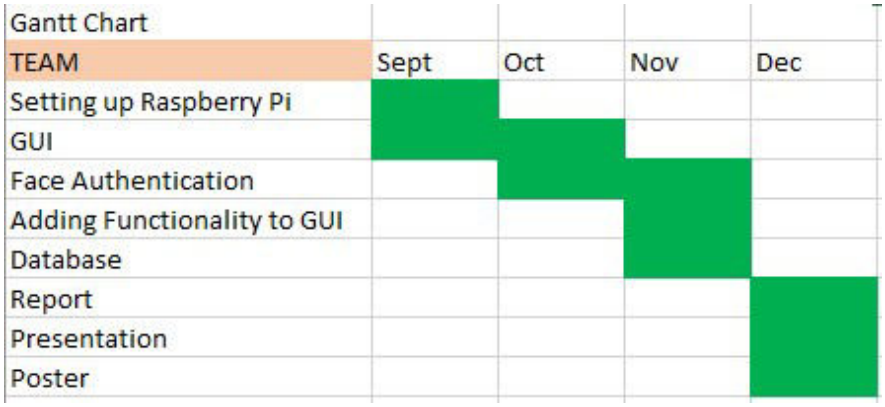


Figure 22: Gantt chart showing team's project timeline

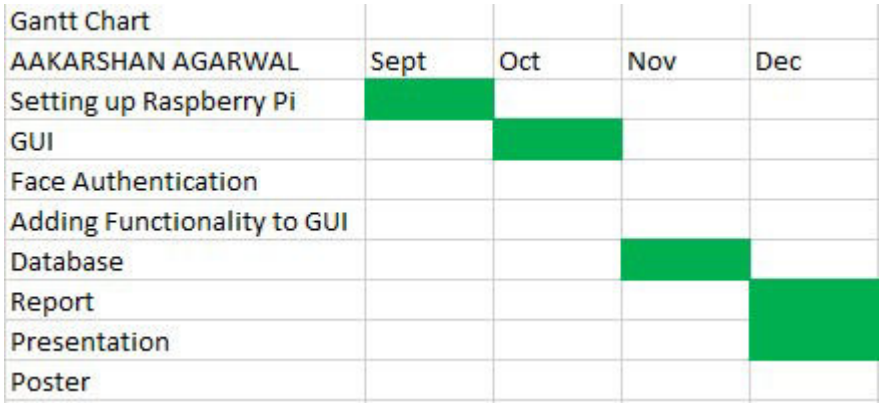


Figure 23: Gantt chart showing Aakarsha's project timeline

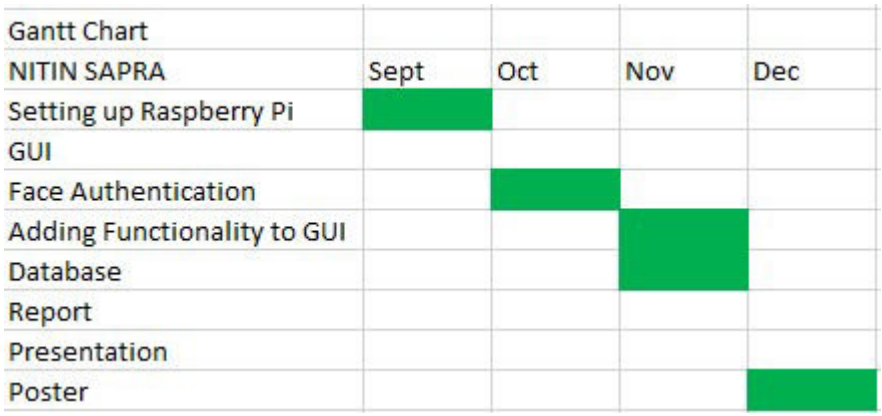


Figure 24: Gantt chart showing Nitin's project timeline

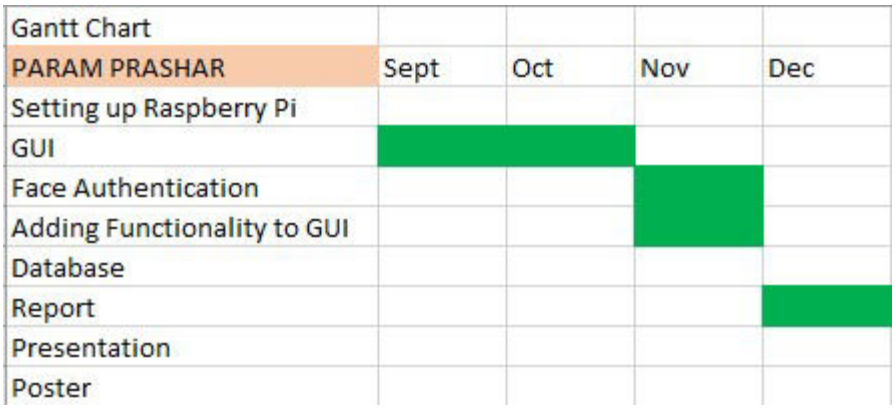


Figure 25: Gantt chart showing Param's project timeline

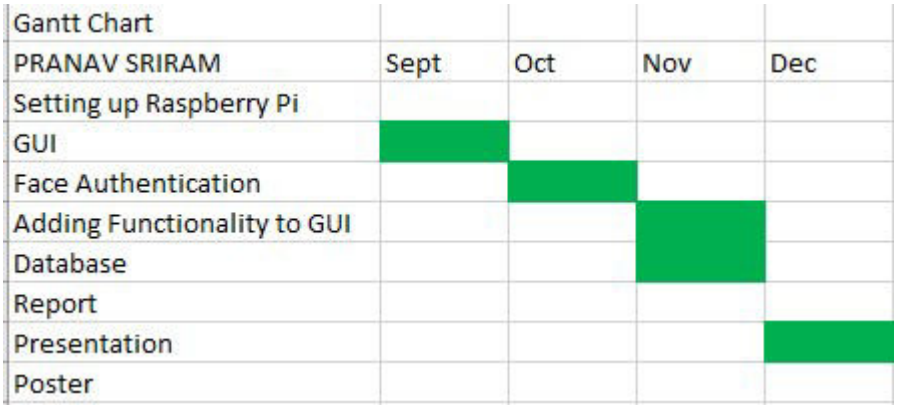


Figure 26: Gantt chart showing Pranav's project timeline

6.2 PEER ASSESSMENT MATRIX

The following table represents the peer assessment done by the team members

		Evaluation of			
		Aakarshan Agarwal	Nitin Sapra	Param Prashar	Pranav Sriram
Evaluation By	Aakarshan Agarwal	-	5	5	5
	Nitin Sapra	5	-	5	5
	Param Prashar	5	5	-	5
	Pranav Sriram	5	5	5	-

Table 4 : Peer Assessment matrix

REFERENCES

1. https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08?gi=1ddeb177c6d8
2. https://blog.paperspace.com/implementing-gradient-boosting-regression-python/
3. https://scikit-learn.org/stable/modules/model_evaluation.html
4. https://towardsdatascience.com/understanding-random-forest-58381e0602d2
5. https://monkeylearn.com/blog/introduction-to-support-vector-machines-svm/
6. Concepts, Tools, and Techniques to Build Intelligent Systems, Aurelien Geron ISBN: 9789352139057
7. https://www.analyticsvidhya.com/blog/2018/06/comprehensive-guide-for-ensemble-models/
8. https://blog.statsbot.co/ensemble-learning-d1dcd548e936
9. https://machinelearningmastery.com/ensemble-methods-for-deep-learning-neural-networks/
10. http://engineering.electrical-equipment.org/electrical-distribution/electric-load-forecasting-advantages-challenges.html
11. https://www.sciencedirect.com/science/article/pii/S0169207015001508

APPENDIX I

Original Metric Outputs

```
[Command: python -u 'C:\Users\Param Prashar\Desktop\Capstone Final Files\LinearReg.py']
Success
Training Accuracy = 0.750, Test Accuracy = 0.690, RMSE (train) = 0.087, RMSE (test) = 0.103
Linear Regression
exp variance err 0.6911093562130763
max error 0.23434762194626568
mae 0.08664719274441536
mse 0.010690458590657969
mean sq log error 0.00022445118341963896
median absolute error 0.06740590675915303
r2 0.6901615108811969
Success!
[Finished in 1.693s]
```

Fig 27 Linear Regression

```
[Command: python -u 'C:\Users\Param Prashar\Desktop\Capstone Final Files\SVM.py']
Success
Training Accuracy = 0.769, Test Accuracy = 0.726, RMSE (train) = 0.083, RMSE (test) = 0.097
SVM
exp variance err 0.7266234161078209
max error 0.19041082853440017
mae 0.08447395807666894
mse 0.009447377764089403
mean sq log error 0.0001999352512615147
median absolute error 0.072345370229848
r2 0.7261893652421998
Success!
[Finished in 1.667s]
```

Fig 28 SVM

```
[Command: python -u 'C:\Users\Param Prashar\Desktop\Capstone Final Files\KNN.py']
Success
Training Accuracy = 0.849, Test Accuracy = 0.813, RMSE (train) = 0.067, RMSE (test) = 0.080
KNN
exp variance err 0.8134579065745691
max error 0.22135864813887896
mae 0.06512917288334664
mse 0.0064607743879589105
mean sq log error 0.00013630861430812658
median absolute error 0.04659169824493414
r2 0.8127492326052363
Success!
[Finished in 1.306s]
```

Fig 29 KNN

```
[Command: python -u 'C:\Users\Param Prashar\Desktop\Capstone Final Files\GradBoost.py']
Success
Training Accuracy = 0.732, Test Accuracy = 0.598, RMSE (train) = 0.090, RMSE (test) = 0.118
Grad Boost
exp variance err 0.6017441237909853
max error 0.2885114910799267
mae 0.09211976428450729
mse 0.01386044634586214
mean sq log error 0.00029195289259090993
median absolute error 0.064159100881934
r2 0.5982866667603035
Success!
[Finished in 1.84s]
```

Fig 30 Gradient boost

```
[Command: python -u 'C:\Users\Param Prashar\Desktop\Capstone Final Files\RandomF.py']
Success
Training Accuracy = 0.968, Test Accuracy = 0.846, RMSE (train) = 0.031, RMSE (test) = 0.073
Random Forest Metrics on Substation Electricity Data
Today's date: 2020-08-08
exp variance err 0.8472739530800768
max error 0.1554127704004724
mae 0.06245151849432849
mse 0.0053087362696602565
mean sq log error 0.00011205451967397156
median absolute error 0.05941045105751641
r2 0.8461384223162227
Success!
[Finished in 3.737s]
```

Fig 31 Random Forest