ijRASET

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

www.ijraset.com

Call: ◎08813907089 | E-mail ID: ijraset@gmail.com

Significant Features and Identification of Heart Disease Prediction by using Data Mining Techniques

M. Deepika[1], P. Jayasimman[2], G. Sanjay[3], M. Sivaprakash[4], T. P. Vignesh[5]
[1]Assistant Professor, Computer Science and Engineering, Nandha Engineering College, Erode
[2, 3, 4, 5]Student, Computer Science and Engineering, Nandha Engineering College, Erode

Abstract: In day to day life, there are various factors that affect the mortal heart. Numerous problems are being at a rapid-fire pace and novel heart conditions are fleetly identified. In this stressful of world, Heart, being an essential organ in the body pumps blood through the body for blood rotation essential and its health is to be conserved for a healthy living. The main provocation of doing this design is to provide a heart complaint prediction model for the prediction of circumstances of heart complaints. Further, this exploration study is aimed towards relating the algorithms to relate the possibility of heart complaint in a case. The identification of possibility of heart complaints in a person is complicated process for medical interpreters because it takes times of experience and violent medical tests need to be conducted. In this study, two data mining algorithms such as KNN and SVM classification are addressed and used to develop the prediction system in order to dissect as well as prognosticate the possibility of heart complaint. The main idea of the significant exploration work is to identify algorithms suitable to provide maximum accuracy when classification of normal/abnormal person is carried out. Therefore prevention of loss of lives at an earlier stage is now possible. It is sure that the above algorithms perform better when compared to other algorithms for heart complaint prediction. The design is designed using Python 3.7.
Keywords: Data mining, Prediction model Classification algorithms, Feature selection, Heart disease prediction

I. INTRODUCTION

There may also be several inheritable factors through which a heart complaint type is passed down from generations. According to World Health Organization, every time more than twelve million deaths are being worldwide due to colorful types of heart conditions which is also known by cardiovascular complaint. The term heart complaint includes various conditions that are different and specifically affect heart and highways of d:a mortal being. Indeed youthful aged people around their 20-30 times of lifetime are getting affected by heart conditions. The increase in possibility of heart complaint among youthful may be due to bad eating habits, restless nature, lack of sleep, depression and multitudinous other factors similar as rotundity, family history, poor diet, high blood pressure, idle geste, high blood cholesterol, family history, smoking and hypertension. The opinion of heart conditions is an important and is the most complicated task in medical field. All the mentioned factors are taken into notable consideration when assaying as well as understanding the cases by croaker through homemade check-ups at regular intervals of time. The heart complaint symptoms greatly depend upon which of discomfort felt by an existent. Some symptoms are not generally linked by the common people. Still, common symptoms include chest pain, breathlessness, and heart pulsations. The chest pain common to numerous types of heart complaint is known as angina, or angina pectoris, and occurs when a part of heart doesn't admit sufficient oxygen. Angina is started by stressful events/physical exertion and typically lasts under ten twinkles.

Heart attacks also do as a result of various types of heart complaint. The sign of the heart attack is analogous to angina except that they do during rest and tend to be much more severe. The symptoms of the heart attack can occasionally act indigestion.

Heartburn and a stomach pang do, as well as a heavy feeling in the chest. Other symptoms of heart attack include pain travels through the body, for illustration from casket to the arms, neck, back, tummy, or jaw, flightiness and dizzy sensations, nausea and vomiting and gushing sweating,. Heart failure is an outgrowth of heart complaint, and breathlessness do when the heart becomes too weak to blood circulation.

Some heart conditions do with no symptoms at each, especially in aged grown-ups and individualities with diabetes. The term' natural heart complaint'covers a range of conditions, but the general symptoms include sweating, high situations of fatigue, fast twinkle and breathing, breathlessness, casket pain. Still, these symptoms might not develop until a person is aged than 13 times.

In these types of cases, the opinion becomes an intricate task taking great experience and high skill. A threat of a heart attack or the possibility of the heart complaint if linked beforehand, can help the cases take preventives and take nonsupervisory measures. Lately, the healthcare assiduity has been generating huge quantities of data about cases and their complaint opinion reports are being especially taken for the analysis of heart attacks worldwide. When the data about heart complaint is huge, the machine literacy ways can be enforced for the analysis.

Data Mining is a task of rooting the vital decision making information from a philanthropy of once records for unborn analysis or vaticination. The information may be hidden and isn't identifiable without the use of data mining.

The classification is one data mining fashion through which the unborn outgrowth or prognostications can be made grounded on the literal data that's available. The medical data mining made a possible result to integrate the bracket ways and give motorized training on the dataset that further leads to exploring the retired patterns in the medical data sets which is used for the analysis of the case's unborn state. Therefore, by using medical data booby-trapping it's possible to give perceptivity on a case's history and is suitable to give clinical support through the analysis. For clinical analysis of the cases, these patterns are veritably important essential. In simple English, the medical data mining uses bracket algorithms that are a vital part for relating the possibility of heart attack before the circumstance.

The classification algorithms can be trained and tested to make the prognostications that determine the person's nature of being affected by heart complaint. In this exploration work, the supervised machine learning conception is employed for making the prognostications. A relative analysis of the three data mining bracket algorithms videlicet Random Forest, Decision Tree and Naïve Bayes are used to make prognostications. The analysis is done at several situations of cross confirmation and several chance of chance split evaluation styles independently.

The StatLog dataset from UCI machine learning depository is employed for making heart complaint prognostications in this exploration work. The prognostications are made using the bracket model that's erected from the bracket algorithms when the heart complaint dataset is used for training. This final model can be used for analysis of any types of heart conditions.

II. RELATED WORKS

In this paper (1) the authors stated that prognostic of life for cases with heart failure remains poor. By using data mining techniques, the purpose of this study is to estimate the most important criteria to prognosticate patient survival and to outline cases for estimating their survival chances together with the most suitable fashion for health care. Five hundred and thirty three cases are suffered from cardiac arrest included in the analysis.

They performed classical i) statistical analysis and ii) data mining analysis using substantially Bayesian networks. The mean age of 533 cases was 63 ($\pm$ 17) and sample was composed of 390 (72) men and 143 (28) women. Cardiac arrest was observed at home for 412 (77) cases, in public place for 62 (12) cases and on public trace for 60 (11) cases. The belief network of variables showed that the remaining alive probability after heart failure is directly associated to 5 variables coitus, age, the original cardiac meter, the origin of the heart failure and technical reanimation ways employed.

Data booby-trapping styles help clinicians to prognosticate the survival of cases and also acclimatized their practices. This work was carried out for each medical procedure and medical problem and it came possible to make a decision tree fleetly with data of a service. The comparison among classic analysis and data mining analysis showed us the donation of data mining system for sorting variables and conclude on significance or the impact of data and variables on the criterion of study. The main limit of system is knowledge accession and necessity to obtain sufficient data to yield an applicable model.

Cardiac arrest is denoted as a robotic unrecoverable arrest of the general rotation by the cardiac inefficacity. It's honored with the absence of femoral palpitation for further than 5 seconds. Without reanimation, the cardiac arrest leads to the unforeseen cardiac death. The public health impact of the unforeseen cardiac death is heavy since survival rate is estimated at between one and twenty for cardiac arrest cases. This represents to deaths the time in United States and to deaths in France. The profiles of the cases are now well known since it generally concern men from about 40 to 75 times.

Hospitalization should be optimal and fast. According to type of cardiac attacks, procedure of supposition of the responsibility varies and some studies show interest of ways over others, according to the cause of the cardiac arrest. Heart complaint is number one cause of death in the U.S. According to the American Heart Association, an estimated people in the U.S. have a coronary attack each time.

Ninety five percent of unforeseen cardiac arrest patients die before reaching sanitarium and heart complaint further lives each time than following six leading causes of death combined (habitual lower respiratory conditions, cancer, accidents, influenza, diabetes mellitus, and pneumonia).

Nearly people in the U.S. die from heart complaint each time are below the age of 65. These data show that the interest to prognosticate the death threat after heart failure and the requirement to dissect events that passed during care to give the prognostic information.

Classic statistical analyses have formerly been done and gave some information related with epidemiology of the heart failure and failure causes. This paper presented the use of the probability in statistical approach to outline heart failure in case samples and prognosticate the impact of events in the care process.

They concluded that it seems that the use of Bayesian network in medical analysis is useful to explore data and to find hidden connections between events and or characteristics of the samples. It is the first approach to agitate suppositions between clinicians and statistical experts. The main limit of the tools is the necessity to have sufficient data to find chronicity in the connections.

In this paper (2) the authors stated that after a decade of abecedarian interdisciplinary exploration in the machine literacy, spadework in this field has been done; 1990s should see wide exploitation of the knowledge discovery as an aid to assembling knowledge bases. The contributors to AAAI Press book Knowledge Discovery in the Databases were excited at implicit benefits of this exploration. The editors hope that some of the excitement will communicate itself to the AI Magazine compendiums of this composition. It has been estimated that quantum of information in the world doubles every twenty months. The size and databases count presumably increased indeed briskly. In 1989, total number of databases in the world is estimated at 5 million, although utmost of them are tiny DBASE III databases. The robotization of business conditioning produced an ever adding sluice of the data because indeed simple deals, similar as a telephone call, use of the credit card, or the medical test, are generally recorded in computers. Scientific as well as government databases are also fleetly growing. The National Aeronautics and Space Administration had much further data than it dissect. Earth observation satellites, planned for 1990s, are anticipated to induce the one terabyte (1015 bytes) of data daily — further than all previous operations combined. At a rate of 1 picture each alternate, it would take the person several times (working nights/weekends) just to look at the film land generated in each day. In biology, federally funded Human Genome design will store the thousands of bytes for each of several billion inheritable bases.

Closer to everyday lives, the 1990 U.S. tale data of a million million bytes render the patterns that in retired ways describe the cultures and mores of moment's United States. What are we supposed to do with this deluge of theraw data? Easily, little of it'll ever be seen by the mortal eyes.

Still, it'll have to be anatomized by the computers, if it can be understood at all. Although simple statistical ways for the data analysis were developed long ago, advanced ways for the intelligent data analysis are not yet mature.

As a result, there's a growing gap between the data generation and data understanding. At the same time, there's a growing consummation/anticipation that data, intelligently anatomized and are presented, will be the precious resource to be used for the competitive advantage. The computer wisdom community is responding to the scientific and practical challenges presented by need to get the knowledge adrift in the deluge of data.

In assessing the eventuality of the AI technologies, Michie (1990), the leading European expert on machine literacy, prognosticated that "the coming area that is going to explode is machine literacy tools usage as an element of large-scale data analysis." The recent National Science Foundation factory on the database exploration future ranked data mining among the most promising exploration motifs for 1990s. Some exploration styles are formerly well enough developed to made part of software commercially available. Several expert system shells also use variations of ID3 to convert rules from exemplifications. Other systems use neural net, inductive or inheritable literacy approaches for discovering patterns in particular computer databases. Numerous forward-looking companies use these and other tools to dissect their databases to intrigue and useful patterns.

The American Airlines finds its frequent leaflet database for finding its better guests, targeting them for the specific marketing elevations. Farm Journal analyzes its subscribers' databases and uses advanced printing technology for custom-figuring hundreds of editions acclimatized to specific groups. Several banks, use patterns discovered in the past loan and credit histories, deduced better loan provisioning. The General Motors also use a database of machine trouble reports to decide the individual expert systems for their colorful models. Packaged-goods manufacturers search supermarket scanner data to measure goods of their elevations and to look for the shopping patterns. A combination of business exploration interests has produced adding demands for, and increased exertion to give, tools and ways to discover in databases. This book is the first to give together leading- edge exploration from around world on these contents. It spans various different approaches to discovery, including inductive literacy, Bayesian statistics, knowledge accession for expert systems, semantic query optimization information proposition, and fuzzy sets.

The book is aimed at those interested in computer wisdom and operation of data, to inform and inspire farther exploration operations. It will be of particular interest for professionals working in databases and operation information systems and to persons applying machine literacy to real-world problems.

III. METHODOLOGY

A. Big Data Analytical Model

Big data analytics support conception of artificial intelligence at heart of numerous new digital health platforms and the perfection health tools. Immaculately, application of the big data logical tools in cardiovascular care will restate into better care and issues at the lower cost.

The eventuality for important prophetic models is a charming operation of the big data analytics. Historically, the prediction models have reckoned on a limited number of specified variables manually entered to estimate the threat score. Similar models generally warrant perfection when they performs nicely well at population position, but not at individual patient position. And despite actuality of dozens of threat models related to the cardiovascular conditions, many are employed to make the remedial opinions.

The big data analytics yield more important analysis of issues ranging from the mortality to case-reported issues to resource application, and so could be more clinically applicable. Machine literacy, for illustration, evaluated the patterns associated with an outgrowth directly from data, rather than from a pre-denoted set of variables.

A full range of associations and relations among data were assessed. Traditional statistical models were being done; machine literacy used a training process whereby model is iteratively given varied data sets to explore the numerous combinations of prophetic features for optimizing analysis. Phenol-mapping, or deep phenol-typing, is one another promising operation in big data. Present complaint groups or phenotypes, are squishy and miscellaneous big data analytics identifies analogous case clusters, create multiple phenotypes within every complaint reality. In proposition, further refined the phenol-mapping of complaint, countries and circles help inform more customized-health opinions

Big data support the combination of multiple data sources from large case populations to estimate the implicit benefits of the curatives similar as ICD's for separate cases. Indeed, the big data are central to the victory of perfection health, given growing interest in incorporating the data, which extensively increases size and complexity of the datasets. Similar datasets bear the advanced logical platforms that are the emblems of big data analytics.

The Big data analysis guide programs to address the certain case member by particular interventions. The success of policy is critically dependent on quality of the underpinning exploration and quality (effectiveness) of the interventions. For various interventions (for case in social/ internal health sphere) widely accepted styles to validate success were still lacking. There are several challenges result regarding the Big Data and populate heart complaint similar as

- Data protection regulations make it delicate to dissect data from various heart complaint providers and services in combination
- Significant parts of population health records are unshaped heart complaint textbook.
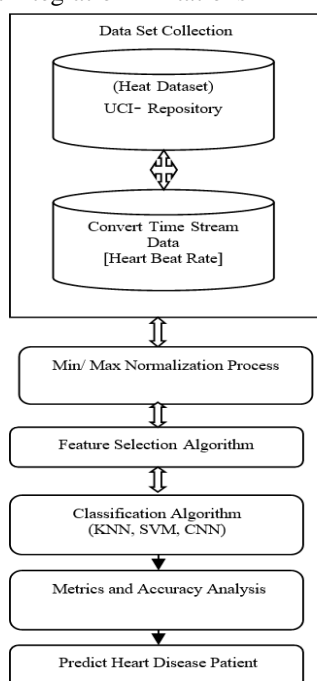- There are data quality, interoperability and data integration limitations

Fig 3.1 Architecture Diagram Of Proposed System

B. Normalization Model

Data transformation such as Normalization is a data preprocessing tool used in the data mining systems. An attribute of the dataset is normalized by scaling the values so that they fall within the small-specified range, like 0.0 to 1.0. Normalization is particularly useful for the classification algorithms which involve neural networks, or distance measurements like nearest neighbor classification and clustering. There are many other methods for data normalization which include min-max normalization, normalization by decimal scaling and z-score normalization.

Min-max normalization performs well in a linear transformation on original heart dataset. Min-max normalization maps the value d of P to d′ in the range [new_min (p), new_max (p)].

Min max normalization preserves relationship among the actual heart dataset values. The table 4.1 describes the sample normalized heat disease dataset model details which show the following details.

Attribute	Original Values	Normalized dataset
age	70.0	35.0
chest pain type	1.0	0.0
resting blood pressure	130.0	140.0
maximum heart rate achieved	109.0	79.0
exercise induced angina	0.0	1.0

Table 3.1 Normalized Heart Dataset

C. Greedy Feature Extraction Model

Feature selection is one of the dimension reduction methods that have been used to allow the better understanding of the data and ameliorate performance of other literacy tasks. Although the applicable features selection has been considerably studied in the supervised literacy, point selection with absence of class markers is still the grueling task.

This paper developed a new system in unsupervised point selection that efficiently selects features in the greedy manner. The paper first defined an effective criterion to unsupervised point selection that measured the reconstruction error of data matrix grounded on the named features subset. The paper also presented a novel algorithm to greedily minimize the reconstruction error grounded on features named so far. The greedy algorithm is grounded on the effective recursive formula to calculate the reconstruction error.

The greedy algorithm selects replication most representative point among remaining available features, and eliminates the effect of named features from given data matrix. This step makes it less likely for algorithm to elect features which are analogous to preliminarily named attributes that consequently reduces the redundancy between named features. Moreover, the use of the recursive criterion made the algorithm computationally doable as well as memory effective compared to the state of the art methods for unsupervised (both forward and backward) point selection.

D. Classification Analytical Model

Machine Literacy indicates how computers learn or ameliorate their performance using the data. System programs to generally plant to descry styles and prove specialized predicated opinions on the data. Machine knowledge is the fast growing discipline. Then, using classic problems in machine literacy that largely is affiliated to data mining.

1) Supervised Bracket Literacy

Supervised bracket literacy model correspond of all the data is labeled and the algorithm will learn to prognosticate the affair from training dataset.
E.g. SVM.

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue IV Apr 2023- Available at www.ijraset.com

2) Unsupervised Bracket Learning

Unsupervised bracket literacy is used to cluster grounded algorithm. In this session the entire information is mot included as well as algorithm founds to essential structure from the given input dataset.

E.g. K- means, KNN. Neural Networks

3) Semi-supervised bracket Learning

Semi-supervised literacy is the combination of supervised literacy as well as unsupervised literacy. In Semi-supervised literacy some data are labeled and some data are not labeled. In this approach here, labeled training dataset were used to learn class models and then unlabelled training dataset are used to define boundaries between classes.

IV. RESULTS AND DISCUSSIONS

A. Dataset Description

Preparing the database-for carrying the result, this paper used Heart patient data sets from ILPD (Indian Heart Case) Data Set (table 4.1).

Attributes Type	Description	Gender Categorical
age	age given in years	Real number
Sex	sex (Value 1 : male; Value 0 : female)	String
Cp	chest pain type(1: typical angina ; 2: atypical angina 3: non-anginal pain ;4: asymptomatic)	Real number
Trestbps	resting blood pressure (in mm Hg on admission to the hospital)	Real number
Chol	Cholestoral(Serum cholestoral) in mg/dl	Real number
Fbs	Fasting blood sugar in mg/dl (>120) Value 1 = true; Value 0 = false)	Real number
Restecg	Resting electrocardiographic results	Real number
Thalach	Heart rate achieved at maximum	Integer
Exang	Exercise induced angina (Value 1 : yes; Value 0 : no)	Integer
Oldpeak	ST depression originated by exercise relative to rest	Integer
Slope	Slope of the peak exercise ST segment (Value 1: upsloping ; Value 2: flat ; Value 3: downsloping)	Integer
Ca	Major vessels (0-3) colored by flouroscopy	Integer
Thal	Result of thalium stress test (Value 3 = normal; Value 6 = fixed defect; Value 7 = reversable defect)	Integer
Num	status of heart disease (angiographic status) Value 0: < 50% diameter narrowing Value 1: > 50% diameter narrowing	Binary

Table 4.1 Dataset Attribute

Table 4.1 describes the attribute type, description and Gender Categorical values.

Fig 4.1 Performance of SVM Classification

Fig 4.2 Performance of KNN Classification

Fig 4.1, Fig 4.2 describes the performance analysis of SVM and KNN classification algorithms respectively.

V. CONCLUSION

The project focused on the SVM classification algorithms for effectively detecting heart disease types. The dataset was taken from UCI repository. Preprocessed like zero values, N/A values and unicode character removal are also carried out here. Important attributes are extracted out for better classification. Confusion matrix is also prepared with accuracy score value calculations. Moreover, KNN classification algorithms and also neural network to effectively detect risk types are carried out. The dataset is taken and preprocessed like unicode removal. Important attributes are extracted out for better classification. Confusion matrix is prepared with the accuracy score calculation. Accuracy prediction is then carried out. Convolutional Neural Network based prediction model was worked out to detect algorithm efficiency. 510 training records and 240 test records are taken out for the convolutional neural network training. There are several directions for future study and research. The current investigation of the classifications is still preliminary.

REFERENCES

[1] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", Studies in Health Technology and Informatics, Vol. 107, No. 2, pp. 1256-1259, 2004.

[2] W.J. Frawley and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview", AI Magazine, Vol. 13, No. 3, pp. 57-70, 1996.

[3] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", Intelligent Computing in Signal Processing and Pattern Recognition, Vol. 345, pp. 721-727, 2006.

[4] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", International Journal of Biological, Biomedical and Medical Sciences, Vol. 3, No. 3, pp. 1-8, 2008.

[5] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", International Journal of Computer Science and Network Security, Vol. 8, No. 8, pp. 1-6, 2008

[6] Shantakumar B. Patil and Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol. 31, No. 4, pp. 642-656, 2009.

[7] Nidhi Singh and Divakar Singh, "Performance Evaluation of K-Means and Hierarchal Clustering in Terms of Accuracy and Running Time", Ph.D Dissertation, Department of Computer Science and Engineering, Barkatullah University Institute of Technology, 2012.

[8] Weiguo, F., Wallace, L., Rich, S., Zhongju, Z.: "Tapping the Power of Text Mining", Communication of the ACM. 49(9), 77-82, 2006.

[9] Jiawei Han M. K, Data Mining Concepts and Techniques, Morgan Kaufmann Publishers, An Imprint of Elsevier, 2006

[10] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998

[11] A comparison of antiarrhythmic-drug therapy with implant-able defibrillators in patients resuscitated from near-fatalventricular arrhythmias. The Antiarrhythmics versusImplantable Defibrillators (AVID) Investigators. N Engl JMed. 1997 Nov 27;337(22):1576-83.

[12] R.Wu, W.Peters, M.W.Morgan, "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal of Healthcare Information Management. 16(4), pp. 50-55, 2002.

[13] Mary K.Obenshain, "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, vol. 25, no.8, pp. 690–695, Aug. 2004.

[14] G.Camps-Valls, L.Gomez-Chova, J.Calpe-Maravilla, J.D.MartinGuerrero, E.Soria-Olivas, L.Alonso-Chorda, J.Moreno, "Robust support vector method for hyperspectral data classification and knowledge discovery." Trans. Geosci. Rem. Sens. vol.42, no.7, pp.1530–1542, July.2004.

[15] Obenshain, M.K: "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, 25(8), 690–695, 2004.

[16] Wu, R., Peters, W., Morgan, M.W.: "The Next Generation Clinical Decision Support: Linking Evidence to Best Practice", Journal Healthcare Information Management. 16(4), 50-55, 2002.

[17] Charly, K.: "Data Mining for the Enterprise", 31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer, 7, 295-304, 1998.

[18] Blake, C.L., Mertz, C.J.: "UCI Machine Learning Databases", http://mlearn.ics.uci.edu/databases/heart-disease/, 2004.

[19] Mohd, H., Mohamed, S. H. S.: "Acceptance Model of Electronic Medical Record", Journal of Advancing Information and Management Studies. 2(1), 75-92, 2005.

INTERNATIONAL JOURNAL
FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 ◯ (24*7 Support on Whatsapp)