Early Prediction and Prevention of Lifestyle Diseases

Sakshi Gaur, Sarvesh Sharma and Ayush Tripathi

June 4, 2021

Early Prediction and Prevention of Lifestyle Diseases

Sakshi Gaur
Student-UG
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
sgaur6554@gmail.com

Sarvesh Sharma
Student-UG
School of Computer Science and Engineering
Galgotias University
Greater Noida, India
sarveshsharma4466@gmail.com

Ayush Tripathi
Student-UG
School of Computer Science and Engineering
Galgotias Universtiy
Greater Noida, India
ayushthebetter1@gmail.com

Abstract- **Lifestyle diseases is common among the population today not only in India but also in almost every country. Lifestyle diseases are caused because of the habits that we have in day to day basis. The way one lives his life is the major cause of it. It includes heart disease, hypertension etc. which all may heard of. In our life also, one also come across atleast one person who is either suffering from such diseases or the diseases became the reaason of his death. We also came across many such people who died because they are not aware of their disease and left with no appropriate time for the treatment. That is why, we decided to develop the model which will analyse the data entered by the user and will give the predictions of the diseases which he or she may have chances to suffer from. This not only give the predictions but also gives you the preventive measures that are required to stay safe from the very common lifestyle diseases as well as in case of mild symptoms it provides you with the management techniques also. This project aware the person about his health so that he will have the treatment well in time if required and will save the lives of many people. This project covers three main aspects which are prediction, prevention and management of lifestyle diseases.**

I. INTRODUCTION

Today, people do not have time for the regular checkup. They are so much busy in their works that they rarely have time for their own health. But the thing is that, they can do the analysis if the appropritae application can provide them the overall health status of the person. This is because they need not to give the time separately for this, rather they can just utilise the time for example the time of travelling, etc. The only thing required is that the amrt phones which almost everyone have with them in this century. So, this can be considered as the portable health checker which everyone can use easily just through a mobile application.

In this, we have decided to give the sign up page where the user can sign up using his name, id and password.

Then further modules will have the diseases portion.

Though the wholehealth checker is somewhat more difficult task, so initially we are adding some of the very common diseases like heart disease, breast cancer, etc. Later on, we will keep adding more and more diseases.

II. PROCESSES INVOLVED

A. EARLY IDENTIFICATION OF THE LIFESTYLE DISEASES

If one observes the lifestyle of a person today, they will find that in most of the cases, people are living unhealthy life.

So, it becomes the concern of utmost importance that one has to identify the diseases he or she is suffering at the earliest. This project serves the above purpose. It monitors the long term activity of the user and will identify the disease on the basis of irregular patterns.

B. PREVENTION OF LIFESTYLE DISEASES

If a person is aware about the health, it becomes quite easy to take the cautions so that it can be prevented. After the identification, the app automatically suggests the preventive measures for the person after analysing the irregular patters in his lifestyle. It gives the measures on the basis of probabilties of the diseases that the person can have.

C. MANAGEMENT OF LIFESTYLE DISEASES

A person himself after analysing his health becomes alert and will try to change the styles which are causing that diseases. This app provides the user with the appropriate management tachniques that a person should adopt to cure that disease. Suppose for the instance if the person is lacking in doing normal workout which leads to stress and is causing hypertensiion to that person, then the app will suggest him some of the workout or exercises plan that he should follow to reduce the stress. It also keeps tracks of the eating habits or the glucose level etc. but the important task is to correctly enter the information by the user.

III. WORK PLAN

- The user has to sign up and fill the data asked. After receiving the data, it will process it and analyse it.
- After analysing it will show the user the probabilities for various lifestyle diseases (in this app, cancer, obesity, type-II diabetes, hypertension, asthma, heart disease.) that the person may have.

- It will alert the user to consult the doctor in case of severe symptoms.
- In case of minimum or no symptoms, it will suggest the appropriate preventive measures to user.
- Later on, the user can also update the data accordingly.

IV. CONCEPT USED

Data Mining is a technique of analyzing the huge amount of data in different aspects to discover the useful information or the knowledge discovery. It combines the concepts of artificial intelligence, statistics, probability, machine learning, deep learning and database system technology. The processes of data collection, selection, cleaning, handling the missing values, transformation, mining, evaluation of pattern, and knowledge visualization involved in data mining process. The data is increasing exponentially so as in the case of the health sector. It is also a major data producing sector which is not only heterogeneous but also valuable as it stores the sensitive health information of the person which can even costs the life of a person. The majority of the methods are used to predict, prevent and manage the diseases appropriately and efficiently. The medical diagnosis is subjective and important in other aspect and depends upon the data available and in this case the data entered by the user. Healthcare related data mining is a difficult field as some minor changes may lead to the huge difference in the predictions and will further affect the output. It explores the hidden patterns which further helps in discovery and extracting knowledge in a database to predict diseases that a person may suffer from. We will use both the core models of data mining i.e., descriptive as well as predictive in big data. In case of the descriptive data analysis, it uses user data to identify the patterns in the data and analyze the relationship between various variables and samples. Descriptive models are apriori association rule, data clustering, summarization and visualization. These models are generally developed by using complete data set but we will try to reduce the number of variables or samples required to predict the output which increases its performance as well as the efficiency.

While in case of predictive data analysis, it uses historical data and current data for predicting the probabilities of the future lifestyle diseases or used for diagnosing and curing the diseases as well. But in case of severe symptoms, it will always suggest the user to consult the doctor as soon as possible. Further enhancement of the model can include the nearby hospitals or the clinics available using the google maps. This can be done by several techniques like the Dijkstra's algorithm. CART Decision trees, artificial neural network (ANN), random forecasting and the regression (linear, logistic and ridge) are the commonly used predictive data models.

A Cluster is a collection of the similar objects which are unsupervised. It doesn't have pre-defined categories. K-means, kernel K-means, Gaussian mixture models, clustering, K-nearest neighbor are few examples of algorithms for clustering. The apriori association rule is a data mining process to find the frequent relationships or patterns or associations between variables or sample in a database. As for example, it establishes the rule that if a person is suffering from the hypertension then what are the probable chances of Cardiovascular diseases in that person. These rules are used to analyze and predict the behaviors in the given situation. Support and confidence can also be used. Support indicates that how frequently the item is occurring in the given database and confidence denotes the number of times of true statements' occurrence. The three type of association rules are Positive rule mining (classical rule), negative rule mining and Interestingness measures. Classical rule is based on the frequent item sets in a positive relationship. The examples of this rule mining are Apriori algorithm and FP-growth algorithm.

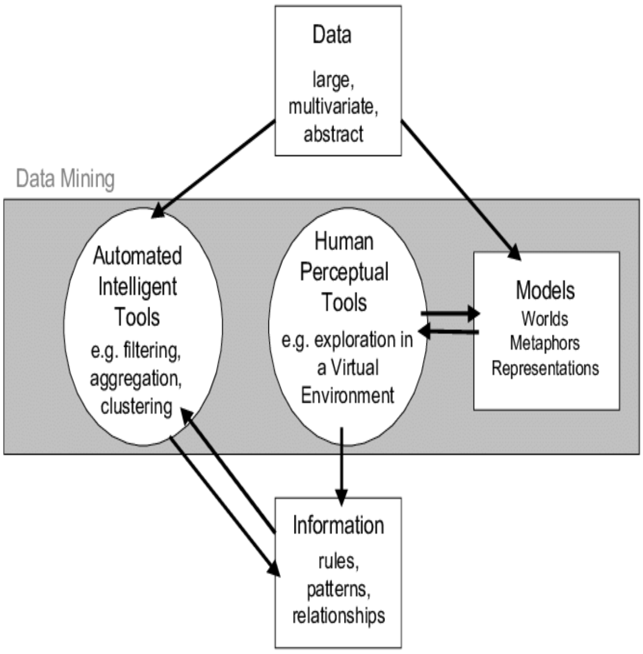
Negative rule mining includes the infrequent item sets in the given database.

Interestingness measures, also known as the constraint-based association rule, applied due to the interest of the users. These can be knowledge based or data constraints. Decision Tree is a structure that consists of the root node, branches and the leaf nodes. It can handle both types of data, that is, numerical and categorical data.

Decision tree gives the presentation of the gathered knowledge. Transformation of leaf nodes to a set of rules by mapping from the root node to the leaf nodes one by one is easy. The result of the decision trees is observed to be highly accurate in case of classification which makes it more reliable and an effective decision-making technique which can suitably be used for medical diagnostics. For more than 20 years, researchers are using these decision trees in medical and health care applications. Some of the instances of the classical decision tree algorithms are CART, ID5R., SADT, and OCI. Hybrid approach is developed to reduce the drawback of the decision trees a by combining the decision trees with the artificial neural networks. All historical approaches were tested on the medical and health domain and some problems were observed in the size, sensitivity and accuracy. This gave momentum to the new approaches as they are able to overcome the drawbacks of the previous approaches. One such example of this is AREX algorithm which is an extended version of decision trees.

Numerical outcome of the model can be well predicted by the technique of regression in which the output variable depends on one or more dependent variables. The predictors can be both numerical or categorical variables. A linear regression will predict the probabilities range between 1 and -1 where 1 is the indication of the linear relationship, 0 shows no relationship and the -1 shows inverse relationship between the variables whereas in case of logistic regression, it produces a logistic curve whose values can vary from 0 to 1. Logistic regression is way more similar to the linear regression in terms that they both gives the relationship between the variables. For the construction of the curve of logistic regression one can use the natural logarithm of the odds against the target variable, rather than just calculating the probability. Nonlinear regression indicates that there exists no linear relationship between dependent and independent variable. Multiple linear regression is used to model the linear relationship between a dependent variable and one or more independent variables. Linear Relationships can further be classified as simple or complex, univariate or multivariate. There are various techniques already available for the regression application some of which can be Generalized Linear Models (GLM), Support Vector Machines (SVM), Ordinary sum of squares (ESS), etc. GLM is used for linear modelling, SVM for linear, nonlinear as well as other mining functions and SEE for the ordinary linear relationship between the variables. For analyzing the medical data, regression turns out to be the most important statistical method (especially using the covariance method) as it establishes the relationship between multiple variables which can be SISO (Single Input Single Output), MISO (Multiple Input Single Output). As for an example, if a patient has high blood sugar level and is influenced majorly by the age, sex and weight. The dependent variable is BSL and independent variables are age, sex and weight. Independent and dependent variables can be selected on the basis as the one that can be measured with the high accuracy can be set as independent. Age can be used in a single variable linear regression or multiple variable linear regression or logistic regression. The strength of the relationship can also be calculated using the covariance.

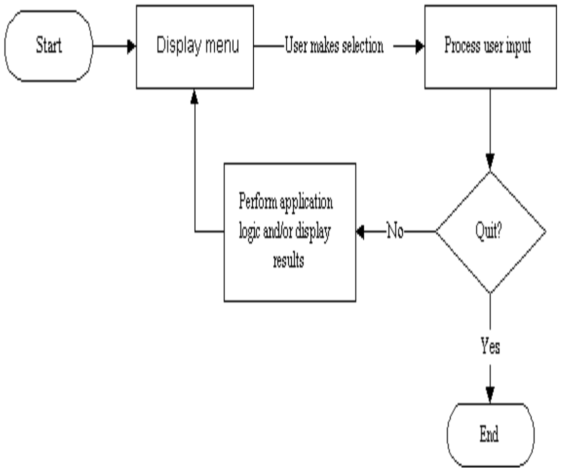
ANN are computational models which are developed parallel to the biological model of the human brain. These are the nonlinear data modelling tools which are quite complex in nature and are also flexible to apply on incomplete, missing and noisy data. When under lying data, relationship is unknown the ANN is the most powerful tool which can be used in the given situation for data modelling. High efficiency and performance can be achieved by combining the ANN with neurofuzzy systems and the genetic algorithms.


Data
large, multivariate, abstract
Data Mining
Automated Intelligent Tools
e.g. filtering, aggregation, clustering
Human Perceptual Tools
e.g. exploration in a Virtual Environment
Models
Worlds
Metaphors
Representations
Information
rules, patterns, relationships

V. ALGORITHMS

A. *For the starting module:*

- For this, we will open the app with its logo and ask the user to sign in.
- In this module, we will be creating a menu driven page with several options for the user where the user will be selecting whether he wants to go to the Prevention Page only or he wants to feed the data for the analysis of prediction of diseases.
- Based on his selection, we will show him the details.
- For an instance, if he will select the Preventive Measures, it will further show him the particular disease for which he wants to search and display accordingly, and if he will select for the analysis then it will ask for the data of the user and based on the algorithms it will further show him up the probabilities of various diseases and the immediate steps to be taken if required otherwise some measures to control.
- For the disease we will be fetching the data from the internet for the preventive measures and update it according to the latest ones.



B. For the Heart Disease:

- In this module, we will be taking into consideration the Heart Disease which is one of the common issues for now.
- Apart from the preventive measures this module will focus on the prediction of heart diseases with the help of various algorithms such as clustering, association rule, etc.
- Clustering is used to predict heart attack in the pre-processed data warehouse.
- Association Rule helps in predicting the chances of heart disease in the person who is already suffering from the other disease.
- Decision Tree helps in determining the heart disease early with maximum possible accuracy through CART algorithm.
- Regression will help in determining the factors that contribute majorly to the Heart Diseases.

C. For the Breast Cancer:

- Breast Cancer is one of the most dangerous disease in females if not treated well in time.
- This module will also provide the preventive measures as well as the chances of the disease in the user.
- Naïve Bayes, Back propagation Neural Network and Decision Tree to predict the survivability rate of breast cancer patients.
- Artificial Neutral Networks, Decision trees and Logistic Regression for the accuracy, sensitivity and specificity of the disease.

D. For Diabetes:

- Diabetes is found in every one of the four persons.
- This will also predict the chance of diabetes along with the preventive measures depending upon the user.
- Clustering is used to predict the likelihood of the disease.
- Association Rule is used to analyse the risk patterns of the Type 2 diabetes.
- Decision tree is used to predict the developing diabetes.
- Regression is used for the predictive analysis of the diabetic treatment.

E. For Hypertension:

- Hypertension is very common among today's generation which is covered by this module.
- There are several techniques which can be used to calculate the risk factors of this disease some of which can be Logistic regression analysis, The CART technique i.e., Classification and Regression decision tree, CHAID (Chi squared Automatic Interaction Detector) and so on.
  This module further provides the user with the preventive measures of the hypertension.

VI. ARCHITECTURE

Input – Data entered by the user.
Processing – Processing is done by using various algorithms of data mining.

Data acquisition is the process of sampling signals that measure real world physical conditions and converting the resulting samples into digital numeric values that can be manipulated by a computer.

Data serialization is the process of converting data objects present in complex data structures into a byte stream for storage, transfer and distribution purposes on physical devices.

Data aggregation is any process whereby data is gathered and expressed in a summary form. When data is aggregated, atomic data rows -- typically gathered from multiple sources are replaced with totals or summary statistics.

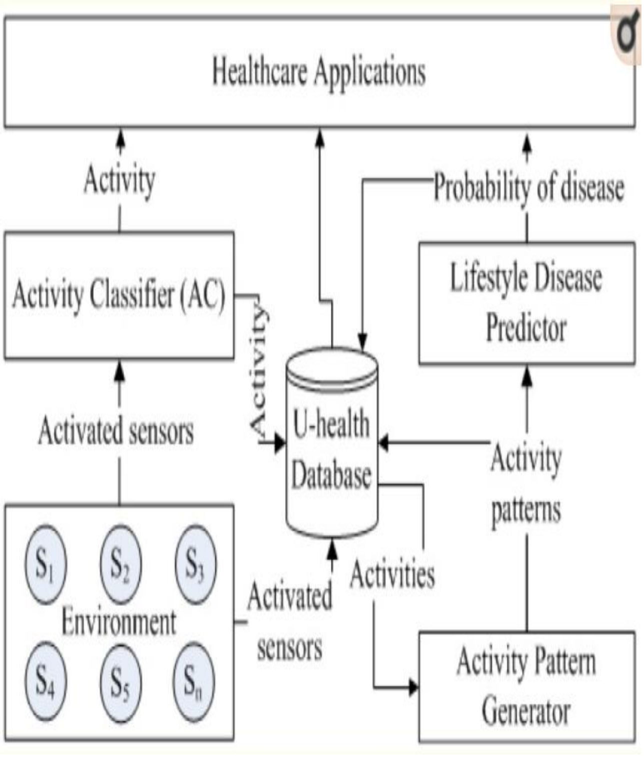
Data analysis is defined as a process of cleaning, transforming, and modelling data to discover useful information for business decision-making. The purpose of Data Analysis is to extract useful information from data and taking the decision based upon the data analysis.

Data mining is a process of extracting patterns and knowledge in huge data sets involving methods of machine learning, statistics, and database systems.

A knowledge representation (KR) is a surrogate, a substitute for the thing itself, used to enable an entity to determine consequences by thinking rather than acting, i.e., by reasoning about the world rather than taking action in it.

Information dissemination is the means by which facts are distributed to the public at large.

Output – Prediction and Preventive measures.



VII. Data Cleaning

Data cleaning is one of the foremost steps to be taken after the extraction of dataset. The dataset that we obtain need not to be appropriate to directly work onto, so cleaning becomes important.

Data cleaning refers to the process of checking the right data types that need to be expected under the desired category of the variables.

With respect to this model, the datatype of age variable must be of integer type, so it must be ensured that all the values entered under the age variable is if integer type.

It holds the similar for the blood pressure that it should be of float type.

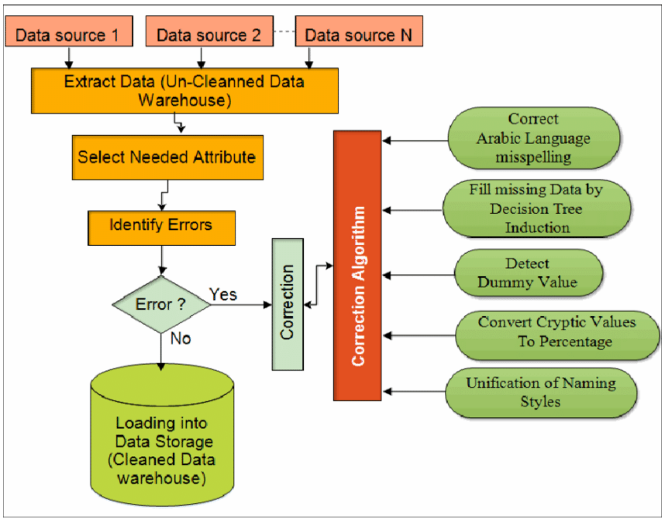
Gender can suitably be considered of string type.

Sleeping hours can be of float type.

So, in this way, variables need to be checked for the datatype.

After this, one should the summary of the variables of how many variables are of categorical type and how many are of numerical type.

Data Deduplication
Quality Check
Data Cleansing
Data Analysis
Data Normalization
Data Standardization


Data source 1
Data source 2
Data source N
Extract Data (Un-Cleanned Data Warehouse)
Select Needed Attribute
Identify Errors
Error ?
Yes
No
Correction
Correction Algorithm
Loading into Data Storage (Cleaned Data warehouse)
Correct Arabic Language misspelling
Fill missing Data by Decision Tree Induction
Detect Dummy Value
Convert Cryptic Values To Percentage
Unification of Naming Styles

VIII. Handling Missing Values

There are fair chances of the values that are not present in the dataset or are missing. So, proper handling of those missing values is important.

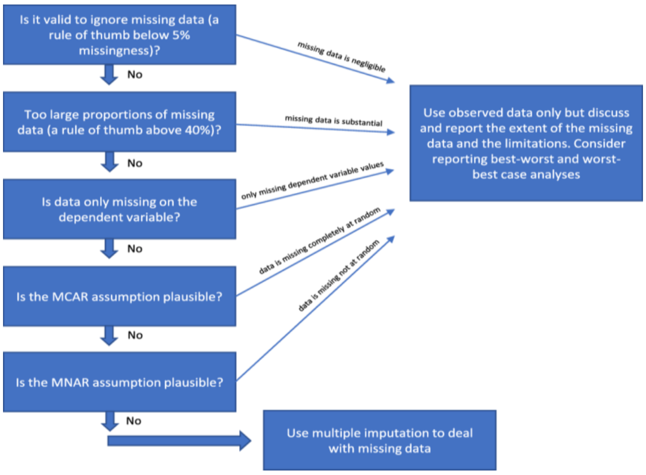
The count of non-null values of each variable gives you the clear idea of how many missing values are present which can be easily calculated with the help of the numpy and panda's library as we are working with the data frames which involve some numerical computations.

One can consider that if any row contains more than half of the variables as missing then complete row can be removed.

If not so, then one can take the help of the statistics to fill those missing values with the mean, median or mode depending upon the difference of all.

Categorical variables can be best filled with the help of mode while for the numerical variables, we can consider the mean and median. If the difference between the mean and median is large, then median is appropriate otherwise mean can be used to fill the missing values.

This can be achieved either by the one-to-one check to the variables or by lambda function to fill the values in one go.

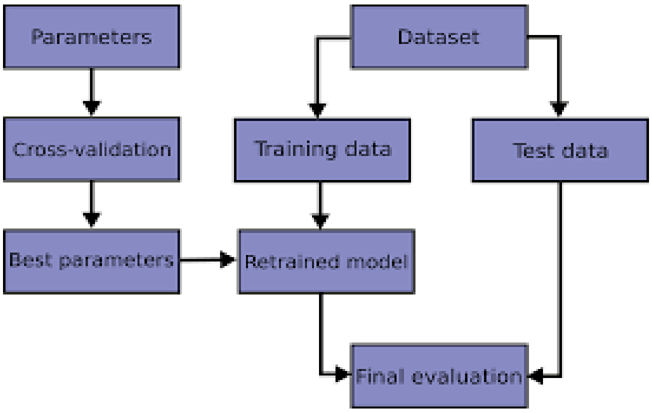

Is it valid to ignore missing data (a rule of thumb below 5% missingness)?
No
Too large proportions of missing data (a rule of thumb above 40%)?
No
Is data only missing on the dependent variable?
No
Is the MCAR assumption plausible?
No
Is the MNAR assumption plausible?
No
missing data is negligible
missing data is substantial
only missing dependent variable values
data is missing completely at random
data is missing not at random
Use observed data only but discuss and report the extent of the missing data and the limitations. Consider reporting best-worst and worst-best case analyses
Use multiple imputation to deal with missing data

IX. Validation

The model can be test and verified by the common technique in the machine learning where a certain fraction of the total dataset which is completely different from the training set. This will help to predict the accuracy of the model. If the model fails to achieve the certain accuracy level then the techniques used need to be modified and again the model should be trained for the dataset and the whole process after the determination of the techniques need to be repeated. In case, it becomes able to achieve the minimum accuracy then one another aspect of the model development can be considered which is to minimise the number of variables required to calculate the output.

This can be achieved by measuring the correlation between the variables and to combine the dependent variables. For the correlation measure, various methods can be used such the Pearson's Correlation which generally used to check and give the linear relation between the variables. Next step to reduce the number of variables would be to check the independency of variables, Apriori association can also be used to check the relation among variables.

Combinations of variables can be formed and then checked the accuracy level of the output being predicted by the model. Step by step, one can reach the stage where certain number of variables can be used to predict the output with the desired accuracy. This is the stage of minimum variables requirement. One can count the variables and use to further enhance the model.



X. Impact and Future Scope

It will help the user to overcome the lifestyle diseases which are in themselves a big threat to humans, will reduce the unawareness about the diseases and will help people to remain healthy which is of utmost importance in today's fast-growing world. It will also change the lifestyle of people for the better. It will also give the clarity about the health of a person or his current status.

In future mostly AI and ML is going to be implemented everywhere. Everyone will get so much busy in their work that they will not get enough time to visit the doctor. Exceptionally when they get serious. But will ignore the minor and common disease which will eventually become more serious in future. Like in tuberculosis person in starting starts to cough and only if taken preventive measures they can be submerged at that time.so this app will become more and more common in future. Without visiting to doctors' people will get to know about their symptoms.

XI. Conclusion

The research on the above diseases have shown that we can use the concept of data mining in this project very efficiently.

Artificial Intelligence is highly used in this project. It can provide already with the several techniques which can calculate the disease risk.

The current system covers the common diseases, the plan is to include disease of higher fatality, like various cancers in future, so that early prediction and treatment could be done, and the fatality rate of deadly diseases like cancer decreases, with the economic benefit in long sight as well.

References

1. https://www.semanticscholar.org/paper/A-Proposed-Model-for-Lifestyle-Disease-Prediction-Patil-Lobo/c32b905b45a5b87fc44459707bbb5a143649b950

2. Data Mining concepts and techniques by Jiawei Han.

3. Mining of massive Datasets by Anand Rajaraman and Jeffrey David Ullman.

4. Ravi N., Dandekar N., Mysore P., Littman L. Activity Recognition from Accelerometer Data.

5. Lifestyle medicine edited by Michael Sagner.

6. https://www.kokilabenhospital.com/blog/tips-to-prevent-lifestyle-diseases/

7. http://ebooks.iospress.nl/volumearticle/48542

8. A Guide to Prevention of Lifestyle Diseases Paperback – 30 October 2004 by R. Kumar, M. Kumar

9. https://en.wikipedia.org/wiki/Lifestyle_disease