International Journal of Science and Research (IJSR)
ISSN: 2319-7064
SJIF (2022): 7.942

# Predicting Health Conditions Using Machine Learning Algorithms on Chronic Diseases

**Leela Prasad Gorrepati**

Richmond, VA, USA

**Abstract:** ***Chronic conditions arise due to various reasons like changes in lifestyle, lack of physical activity, ignorance of early symptoms of diseases etc. Many of such like Heart disease, Diabetes, High blood pressure could be avoided, delayed or could be handled in an efficient way. Also, these are related to each other in one way or the other. 2 out of Top 3 chronic diseases such as Heart disease and Diabetes could be handled by cautious approach. Annually huge sum of amount ranging beyond $400 Billion is being spent on the treatments of these chronic diseases, in the year 2022 the amount is equivalent to nations GDP share of 17.3% [1]. Study has shown that about 60% of US population has 1 or more chronic conditions the equivalents to about 133 Million people which is raising year over year leading and it is predicted to reach 171 million by 2030[2][3]. Despite such a huge expenditure on these chronic diseases, the death rate is also high where it accounts to 70 percent each year due to these chronic conditions [4]. Changes in lifestyle, diet, physical activity and few other incorporations in our daily routine could easily bring down healthcare expenditure along with upgrading quality of life. This would bring down the alarming death rate, the intent of this white paper is to predict the risk of heart failure that they possess.***

**Keywords:** Chronic Conditions, Heart failure, Prediction, Diabetes, Machine Learning, Random Forest Algorithm, lifestyle, Diet, Healthcare, Chronic illness

## 1. Introduction

As age increases, changes in health occur due to various factors like changes in lifestyle, Stress at work, reduced physical activity, frequent traveling, time management, and various other factors. There's no straightforward approach or shortcut to deal with Chronic Conditions, and they are often heard as multifactorial due to their complex nature. However, in reality, it's managed by incorporating minor changes in one's lifestyle, which is mainly being ignored due to various factors, such as ignorance, economic conditions, lack of proper guidance, healthcare plans, smoking, and excessive alcohol intake, diet management, lack of physical activity, etc. When ignored, these chronic conditions lead to adverse effects on the human body, with long-term impacts on the functioning of other critical organs such as the heart, blood pressure, Kidneys, nervous system, etc. This leads to widespread passivity among the population, accompanied by obesity, hypertension, elevated cholesterol levels and other health issues. Life expectancy and quality of life are adversely affected when proper care is not employed while dealing with these chronic conditions. Studies[3] have shown that, in the past decade and a half, the population with diabetes has increased twice and is seen significantly even in people below 65 years of age.

The impact of chronic conditions on people in the United States, in terms of healthcare costs and quality of life, is significant. Making small lifestyle changes and taking preventive measures could improve health and help avoid expensive treatments. It's important to understand an individual's health condition and the associated risks to determine the need for these measures. Machine learning algorithms can be used to predict an individual's risk of heart disease by considering various parameters. This white paper focuses on predicting heart disease using a model that can be trained and utilized by healthcare organizations for effective future planning.

## 2. Solution

Our approach involves utilizing predictive data modeling methods to forecast the likelihood of an individual's possibility of heart failure. Predictive data models are statistical procedures designed to anticipate future outcome based on past data and can be termed a supervised learning approach[5]. These models consist of algorithms and fall into two main types[6]: regression models, which forecast numerical values, and classification models, which predict membership in specific classes. The prediction of heart failure is one aspect, and a similar implementation could be applied to the prediction of other chronic conditions as well.

The Random Forest technique[6] is recognized as a collective classifier algorithm, essentially aggregating decision trees[7]. It leverages the advantages of bagging through the random selection of features. In this scenario, random forest will be employed to forecast the likelihood of an individual having heart failure. This approach enables the creation of a model that healthcare professionals can utilize to pinpoint patients vulnerable to heart failure.

We will use research data gathered from different states and regions to construct this model. This data encompasses medical information and laboratory analyses. The following are the steps we will take in this process:

- Collection of Right Data: Identify and gather data from multiple sources.
- Feature Engineering: Prepare and refine the collected data for analysis.
- Visualize the data: Utilize visualizations to analyze the data to extract critical insights.
- Model Training: Construct predictive models using the analyzed data.
- Validate the trained model: Test and confirm the accuracy of the model's predictions.

**a) Collection of Right Data**

- This dataset contains crucial Medical and Laboratory information. The (.csv) file consists of numerous variables, including independent medical predictor variables, and a single target dependent variable labeled HeartDisease. Below are the attributes in the dataset[10]:
- Age: Represents the patient's age in years.
- Sex: Indicates the patient's sex (M for Male, F for Female).
- ChestPainType: Describes the type of chest pain experienced by the patient, categorized as Typical Angina (TA), Atypical Angina (ATA), Non-Anginal Pain (NAP), or Asymptomatic (ASY).
- RestingBP: The patient's resting blood pressure, measured in millimeters of mercury (mm Hg).
- Cholesterol: The serum cholesterol level of the patient, measured in milligrams per deciliter (mm/dl)[12].
- FastingBS: Indicates whether the patient's fasting blood sugar exceeds 120 mg/dl (1 if yes, 0 otherwise).
- RestingECG: Results of the resting electrocardiogram, classified as Normal, having ST-T wave abnormality (ST), or showing probable or definite left ventricular hypertrophy by Estes' criteria (LVH) [13].
- MaxHR: The maximum heart rate achieved by the patient, a numeric value ranging between 60 and 202.
- ExerciseAngina: Denotes whether the patient experiences angina induced by exercise (Y for Yes, N for No).
- Oldpeak: The measurement of ST depression induced by exercise relative to rest, a numeric value.
- ST_Slope: The slope of the peak exercise ST segment, categorized as upsloping (Up), flat (Flat), or downsloping (Down).
- HeartDisease: The output class indicating the presence of heart disease (1 for heart disease, 0 for Normal).

**b) Feature Engineering**

Feature engineering plays a vital role in supervised learning, involving the deliberate selection, manipulation, and transformation of raw data into usable features for modeling. This process comprises five essential steps: creating features, transforming them, extracting relevant information, conducting exploratory data analysis, and performing benchmarking. Each of these stages plays a pivotal role in enhancing the model's predictive power by ensuring that the input data is optimally prepared for analysis. Through these meticulous processes, feature engineering aims to improve the accuracy and efficiency of predictive models, thereby significantly contributing to the advancement of machine learning applications.

Import the Needed Libraries.

Read the source heart dataset as pandas dataframe.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from sklearn.compose import ColumnTransformer
from sklearn.preprocessing import StandardScaler, OneHotEncoder
from sklearn.ensemble import RandomForestClassifier
```

```
df = pd.read_csv('/content/sample_data/heart_condition.csv')
df.head(5)
```

Python is the language chosen for this prediction as it encompasses a rich set of Machine Learning libraries. Below are some of the libraries that are used for our model.

- Pandas: The library serves as a tool for the manipulation and analysis of data.
- NumPy: The library is utilized for working with large multidimensional arrays.
- Matplotlib: The library is utilized for creating visualizations.
- Seaborn: This library serves the purpose of generating sophisticated visual representations.
- Sklearn: Machine Learning Library featuring various algorithms
- Feature Engineering is a crucial step in training the model for the prediction, below are the function used:
- Describe (): To generate descriptive statistics of a DataFrame, one can utilize the `describe()` method provided by pandas in Python. This method offers a summary that includes count, mean, standard deviation, minimum, maximum, and the quartiles of the dataset for numerical columns. For object-type columns, it provides the count, unique, top, and frequency of the top occurrence. This comprehensive analysis provides essential insights into the distribution and central tendencies of the data, facilitating a deeper understanding of its characteristics.
- isnull(): The isnull() function serves the purpose of identifying missing values within a DataFrame or Series. Its output consists of a boolean mask that denotes the presence of null (True) or non-null (False) elements.
- shape (): To obtain the dimensionality of a DataFrame, one can utilize the `.shape` attribute. This attribute returns a tuple indicating the number of rows and columns in the DataFrame, respectively.
- info(): In Pandas, the `info()` method is utilized to offer a concise summary of a DataFrame. This summary includes details such as the number of entries, the number of non-null values, the data type of each column, and the memory usage

```
# Shape Of The Dataset
df.shape
```

```
(918, 12)
```

```
df.describe()
```

| | Age | RestingBP | Cholesterol | FastingBS | MaxHR | Oldpeak | HeartDisease |
|---|---|---|---|---|---|---|---|
| count | 918.000000 | 918.000000 | 918.000000 | 918.000000 | 918.000000 | 918.000000 | 918.000000 |
| mean | 53.510893 | 132.396514 | 198.799564 | 0.233115 | 136.809368 | 0.887364 | 0.553377 |
| std | 9.432617 | 18.514154 | 109.384145 | 0.423046 | 25.460334 | 1.066570 | 0.497414 |
| min | 28.000000 | 0.000000 | 0.000000 | 0.000000 | 60.000000 | -2.600000 | 0.000000 |
| 25% | 47.000000 | 120.000000 | 173.250000 | 0.000000 | 120.000000 | 0.000000 | 0.000000 |
| 50% | 54.000000 | 130.000000 | 223.000000 | 0.000000 | 138.000000 | 0.600000 | 1.000000 |
| 75% | 60.000000 | 140.000000 | 267.000000 | 0.000000 | 156.000000 | 1.500000 | 1.000000 |
| max | 77.000000 | 200.000000 | 603.000000 | 1.000000 | 202.000000 | 6.200000 | 1.000000 |

```
df.isnull().sum()
```

| | 0 |
|---|---|
| Age | 0 |
| Sex | 0 |
| ChestPainType | 0 |
| RestingBP | 0 |
| Cholesterol | 0 |
| FastingBS | 0 |
| RestingECG | 0 |
| MaxHR | 0 |
| ExerciseAngina | 0 |
| Oldpeak | 0 |
| ST_Slope | 0 |
| HeartDisease | 0 |

The dataset should be free of null or missing values, as their presence can significantly impact the model's accuracy under development. However, should null or missing values be present in the dataset, the following commonly utilized methods can be employed to address them:

- Mean Imputation: Appropriate for Normally distributed data.
- Median Imputation: Appropriate for skewed data distribution.
- Mode Imputation: Appropriate for categorical data.

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Age             918 non-null    int64
 1   Sex             918 non-null    object
 2   ChestPainType   918 non-null    object
 3   RestingBP       918 non-null    int64
 4   Cholesterol     918 non-null    int64
 5   FastingBS       918 non-null    int64
 6   RestingECG      918 non-null    object
 7   MaxHR           918 non-null    int64
 8   ExerciseAngina  918 non-null    object
 9   Oldpeak         918 non-null    float64
 10  ST_Slope        918 non-null    object
 11  HeartDisease    918 non-null    int64
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

Analyzing the "Cholesterol" column revealed that 172 entries were marked with a cholesterol level of "0," which is not feasible since cholesterol levels cannot be zero. Recognizing that employing "0" values could undermine the efficacy of the model, it has been decided to substitute these values with the average cholesterol levels of male and female patients respectively, to ensure more accurate and realistic data representation.

```
df['Cholesterol'].value_counts()
```

| Cholesterol | count |
|---|---|
| 0 | 172 |
| 254 | 11 |
| 223 | 10 |
| 220 | 10 |
| 230 | 9 |
| ... | ... |
| 392 | 1 |
| 316 | 1 |
| 153 | 1 |
| 466 | 1 |
| 131 | 1 |

222 rows × 1 columns

The SimpleImputer library is utilized as a Univariate imputer to fill in missing values using straightforward strategies.

```
from sklearn.impute import  SimpleImputer
imp = SimpleImputer(missing_values=0)
df['Cholesterol'] = imp.fit_transform(df[['Cholesterol']])
df['Cholesterol'].value_counts()
```

| Cholesterol | count |
|---|---|
| 244.635389 | 172 |
| 254.000000 | 11 |
| 223.000000 | 10 |
| 220.000000 | 10 |
| 230.000000 | 9 |
| ... | ... |
| 392.000000 | 1 |
| 316.000000 | 1 |
| 153.000000 | 1 |
| 466.000000 | 1 |
| 131.000000 | 1 |

222 rows × 1 columns

**c) Visualize the data**

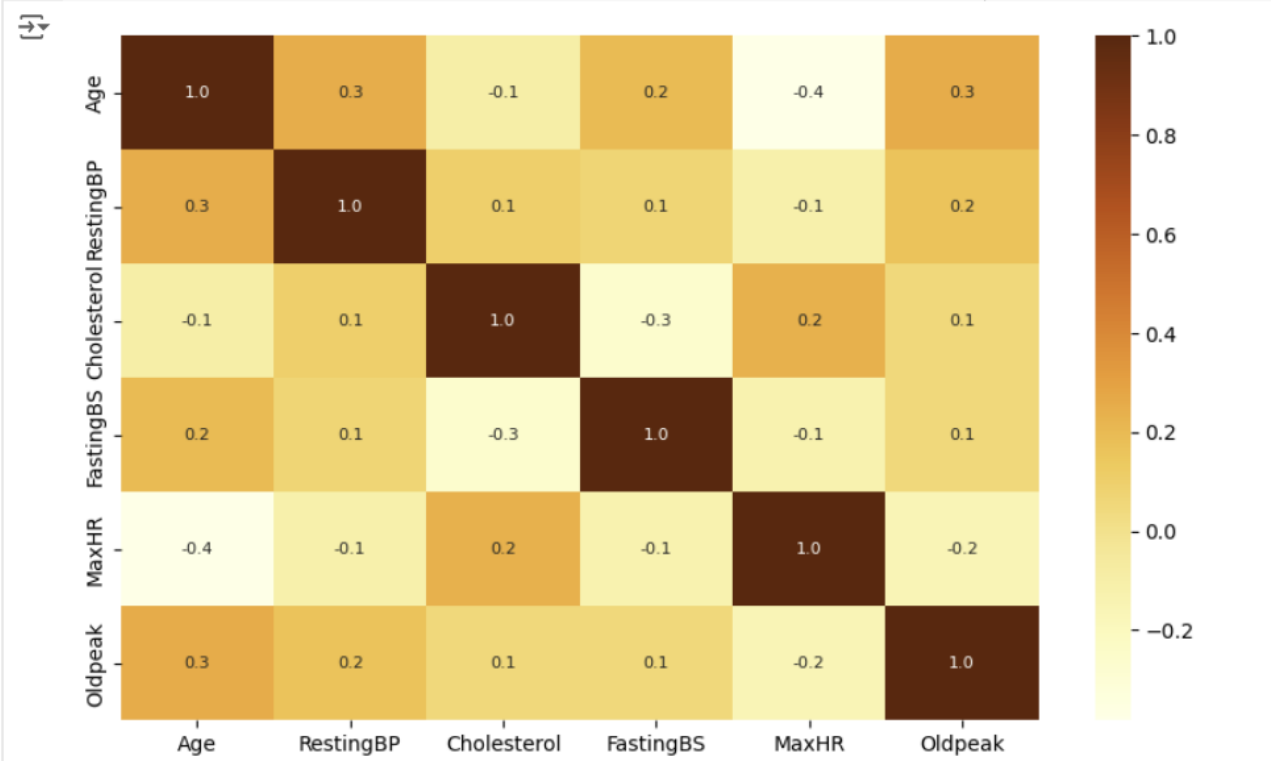
Data visualization is essential in the realm of predictive data analytics, especially in the context of model development. It serves as a critical tool for the elucidation of data patterns, the identification of outliers, the assessment of data distribution, the recognition of skewness, the delineation of trends, the discovery of correlations between variables, and the direction of hypothesis formulation. Here is the Univariate Analysis that can be employed to enhance model construction.

```
# Identify the categorical and numerical columns.
# Exempt target column 'HeartDisease'
categorical_col=df.select_dtypes("string").columns.to_list()

numerical_col=df.columns.to_list()
for col in categorical_col:
    numerical_col.remove(col)
numerical_col.remove("HeartDisease")
```

```
# Correlation numerical features
df_correlational = df[numerical_col].corr()
f, ax = plt.subplots(figsize=(10, 6))

sns.heatmap(df_correlational,cmap='YlOrBr', annot=True, fmt='.1f',annot_kws={'size': 8}, ax=ax)
plt.show()
```




```
# Perform univariate analysis for each numerical_col
for variable in numerical_col:

    plt.figure(figsize=(8, 4))

    plt.subplot(1, 2, 1)
    sns.histplot(data=df, x=variable, kde=True)
    plt.xlabel(variable)
    plt.ylabel("Frequency")
    plt.title("Histogram")

    # Box plot
    plt.subplot(1, 2, 2)
    sns.boxplot(data=df, y=variable)
    plt.ylabel(variable)
    plt.title("Box Plot")

    plt.tight_layout()

    # Presenting plots for visualization
    plt.show()
```

- Seaborn. Heatmap (): Heatmaps are defined as graphical representations of data that utilize colors to visualize the magnitude of matrix values. In these visualizations, brighter colors, predominantly in the reddish spectrum, are employed to denote higher frequencies or activities, whereas darker shades are chosen to signify lower occurrences or activities. Additionally, the term "shading matrix" is synonymous with heatmap. Within the context of Seaborn, a Python visualization library, heatmaps can be efficiently generated using the `seaborn.heatmap()` function.
- Seaborn. boxplot (): This statement summarizes the variability of data values using a visual representation that includes mean, upper and lower quartiles, min and max values, and outliers.

**d) Model Training**

A critical step in model construction involves feature selection. This phase entails revisiting the exploratory data analysis conducted previously and identifying only those essential features that will significantly enhance the model's predictive accuracy. As per the analysis, we will use the attributes– Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST_Slope.

The subsequent phase in the process of constructing a model entails the division of data. This division of the datasets into training and testing subsets is crucial for assessing the performance of the model. The training dataset is utilized for the purpose of training the model, whereas the testing datasets serve to evaluate the model's efficacy.

```
[ ] from sklearn.ensemble import RandomForestClassifier
    model = RandomForestClassifier()
    model.fit(X_train, y_train)
```

```
RandomForestClassifier
RandomForestClassifier()
```

```
[26] model.score(X_test,y_test)
```

```
0.9021739130434783
```

**e) Validate the trained model**

The subsequent phase in the model development process entails evaluating the model's performance by employing the test dataset. This evaluation includes determining the accuracy score of the model that has been trained, followed by testing it on various random datasets to ensure robustness and reliability in its predictions. Here in this below example, with the passed parameters the model has predicted that the person is likely at the risk of heart failure.

This trained model with Random Forest Algorithm predicts if a person is likely to have Heart failure with an accuracy of 90.21%.

```
predict_data = np.array([[79, 1, 2, 138, 287.0, 138, 1, 138, 1,1.5,0]])
# Predict using the trained model
predict_outcome = model.predict(predict_data)
# Display the predict_outcome
print("Prediction Outcome:", predict_outcome)
```

```
Prediction Outcome: [1]
```

**Utilization of the Solution Across Different Organizational Functions**

The Random Forest algorithm, a versatile and powerful ensemble learning method, finds application across a wide range of fields due to its robustness, ease of use, and ability to handle large datasets with high dimensionality. Here are some notable applications:

- **Healthcare and Medicine**: Apart from this Heart failure prediction, in the healthcare sector, Random Forest can be used for the diagnosis of other diseases and medical conditions, such as cancer detection from complex datasets, by identifying patterns and correlations that may not be apparent to human observers. It also aids in predicting disease outbreaks and patient prognosis.
- **Banking and Finance:** Financial institutions could leverage Random Forest for credit scoring, fraud detection, and risk management. By analyzing customer data, the algorithm can predict the likelihood of a customer defaulting on a loan or detect unusual patterns that may indicate fraudulent activity [14].
- **E-commerce and Retail:** Random Forest algorithms help in predicting customer behavior, such as purchase patterns and product preferences. This information is crucial for inventory management, personalized marketing, recommendation systems, and optimizing the customer experience.
- **Manufacturing:** Random Forest can be applied for predictive maintenance, quality control, and supply chain optimization. It can predict machinery failures before they occur, reducing downtime and maintenance costs, and ensuring the quality of manufactured products.
- **Cybersecurity:** In the realm of cybersecurity, Random Forest can be employed to detect and predict security breaches and malware threats. It could analyze patterns in network traffic to identify potential threats, enhancing the security of information systems.
- **Agriculture:** Random Forest could be helpful in precision agriculture, from predicting crop yields to detecting plant diseases and pest infestations. This enables farmers to make informed decisions, leading to increased efficiency and productivity.

## 3. Advantages of the Solution

This solution provides numerous advantages to the global healthcare sector. Below are the main benefits:

- **Controlling heart failure rate**: By effectively managing heart failure, patients enjoy a higher quality of life, allowing them to engage more fully in daily activities.
- **Reduced Hospital Admissions**: Utilizing this predictive model allows healthcare organizations to craft strategies that focus on a patient-centered approach, enhancing overall outcomes to control heart failure rate. This in turn leads to fewer emergency room visits and hospital admissions, significantly reducing healthcare costs and easing the burden on healthcare facilities [15].
- **Extended Life Expectancy**: Utilizing this predictive model allows enables proper management of heart failure, thus potentially extending patients' life expectancy.

- **Increased Healthcare Efficiency**: By reducing the frequency of acute exacerbations and hospitalizations, healthcare resources can be allocated more efficiently, improving care for other patients as well.
- **Enhanced Patient Education and Self-Management**: With this Predictive model, both healthcare companies and the patients could focus on controlling heart failure often involves educating patients about their condition, which empowers them to take an active role in managing their health, leading to better outcomes.
- **Lower Healthcare Costs**: Decreased hospitalization rates and emergency visits directly translate into lower healthcare costs for both patients and healthcare systems, finally it benefits the whole nation.

## 4. Conclusion

In conclusion, the pragmatic application of predictive data analytics is essential for devising cost-effective healthcare strategies aimed at managing chronic conditions such as heart failures. Utilizing data-driven insights enables healthcare organizations to enhance care quality, improve patient outcomes, mitigate prevalence rates, and lower healthcare expenses. This white paper offers a technical viewpoint on the critical importance of prediction in tackling the difficulties associated with chronic conditions like heart failures. It provides recommendations on leveraging data-driven approaches to revolutionize healthcare system delivery.

## References

[1] https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/historical
[2] https://www.rand.org/content/dam/rand/pubs/tools/TL200/TL221/RAND_TL221.pdf
[3] https://www.aha.org/system/files/content/00-10/071204_H4L_FocusonWellness.pdf
[4] https://www.cdcfoundation.org/safeguarding-americans-health#:~:text=Chronic%20diseases%20are%20responsible%20for,our%20nation's%20health%20care%20costs.
[5] https://link.springer.com/article/10.1007/S42979-021-00592-X
[6] http://cmnt.lv/upload-files/ns_51art022_%20CMNT1812-275%20ZHA%20ESMC-E-55%20Corrected%20SH%20ED%20(1).pdf
[7] https://www.jastt.org/index.php/jasttpath/article/view/65
[8] https://scikit-learn.org/1.5/modules/generated/sklearn.impute.SimpleImputer.html
[9] Exploring the Benefits of Comprehensive Evaluations by Health Specialists – Disease prevention. http://diseaseprevention.info/comprehensive-evaluations-by-health-specialists/
[10] Heart disease binary best model data - Data Set Library. https://support.minitab.com/en-us/datasets/predictive-analytics-data-sets/heart-disease-binary-best-model-data/
[11] Exploring the Benefits of Comprehensive Evaluations by Health Specialists – Disease prevention. http://diseaseprevention.info/comprehensive-evaluations-by-health-specialists/
[12] Heart disease binary best model data - Data Set Library. https://support.minitab.com/en-us/datasets/predictive-analytics-data-sets/heart-disease-binary-best-model-data/
[13] Al-Sayed, A., Khayyat, M., Khayyat, M., & Zamzami, N. (2023). Predicting Heart Disease Using Collaborative Clustering and Ensemble Learning Techniques. Applied Sciences, 13(24), 13278.
[14] The hitchhiker's guide to ChatGPT for businesses. https://hitchhiker.kern.ai/glossary
[15] Sunflower Diversified creates tools for health-education program - Great Bend Tribune. https://www.gbtribune.com/news/local-news/sunflower-diversified-creates-tools-for-health-education-program/
[16] Mittal, U., & Panchal, D. (2023). AI-based evaluation system for supply chain vulnerabilities and resilience amidst external shocks: An empirical approach. Reports in Mechanical Engineering, 4(1). https://doi.org/10.31181/rme040122112023m
[17] Sudeep Nagaraj, Dhivya. (2024). Augmented AI in Health Diagnostics: Enhancing Medical Decision Making through Artificial Intelligence. International Journal of Science and Research (IJSR). 13. 1645-1648. 10.21275/SR241022231140.