Revolutionizing Healthcare by Unleashing the Power of Machine Learning in Diagnosis and Treatment

Medini Gupta[1], Sarvesh Tanwar[2], Salil Bharany[3]*, Faisal Binzagr[4],
Hadia Abdelgader Osman[5], Ashraf Osman Ibrahim[6]*, Samsul Ariffin Abdul Karim[7]
Amity Institute of Information Technology, Amity University Noida, India[1, 2]
Independent Researcher, Amritsar 143001, Punjab, India[3]
Department of Computer Science, King Abdulaziz University, P.O. Box 344, Rabigh 21911, Saudi Arabia[4]
Northern Border University, Applied College, Computer Department, Arar, Saudi Arabia[5]
Creative Advanced Machine Intelligence Research Centre-Faculty of Computing and Informatics, Universiti Malaysia Sabah[6, 7]

***Abstract*—Machine learning (ML) is a versatile technology that has the potential to revolutionize various industries. ML can predict future trends in customer expectations that allow organizations to develop new products accordingly. ML is a crucial field of data science that uses different algorithms to predict insights and improve decision-making. The widespread acceptance of ML algorithms ML can provide helpful information using the enormous volume of health data generated regularly. Quicker diagnoses by doctors can be delivered by adopting ML techniques that can bring down medical charges and applying pattern identification algorithms to examine medical images. Every technology brings its challenges; in the same way, ML also has several challenges in healthcare that need to be acknowledged before we witness complete automation in medical diagnosis. People are still forbidden to share their personal information with intermediaries for treatment. Medical record governance is essential to ensure that health records are not missed. Manual diagnosis often goes in the wrong direction, as doctors are also human. Lack of communication between medical workers and patients, considering the insufficient data to diagnose disease, sometimes results in deteriorating health conditions. This paper deals with an introduction to machine learning. These ML algorithms are widely used for health diagnosis, a comparison analysis of literature work that has been done so far, existing challenges of the healthcare system, healthcare industry using machine learning applications, real-life use cases, practical implementation of disease prediction, and conclusion with its future scope.**

***Keywords*—Machine Learning; Health Diagnosis; Supervised Learning; Prediction; Classification**

I. Introduction

Machine learning is a disruptive technology that allows computers to gain knowledge automatically based on the historical data provided. In the last couple of years, this technology has delivered innovative services that have profoundly impacted the human lifestyle. Human beings have the unique ability to learn new things by themselves, depending on their surroundings by healthcare enterprises has fast-tracked medical diagnosis. Training of machine learning algorithms can be done in various ways. ML is the subcategory of artificial intelligence (AI), which mainly deals with developing innovative machinery that holds the capacity to perform work [1] that requires human intelligence. Arthur Samuel is known as the father of AI. ML algorithms are built for computers to get an insight into the outcome based on previous experience. A specific quantity of past data, termed a training data set, is considered. Decisions are being made without the need to be explicitly programmed by developing a mathematical model.

The higher the amount of training data, the greater the performance of the predictive model will be [2]. The historical dataset is given as an input, based on which the ML model builds, and computers predict the result when new data is imparted. Handling complex jobs where human lives can be at stake can be successfully solved by using robots that are programmed using ML algorithms [3]. The precision of the result totally depends on the quantity of the dataset. Transparent ML techniques that develop the proper drug as human safety should always be prioritized [14]. High accuracy can only be achieved by giving machines large amounts of data [18]. Presently, we can recognize various applications of ML such as friend suggestion used by Meta, prediction of traffic by Google Maps, Speech recognition by Google on smart devices, item recommendation on e-commerce websites such as Amazon and Myntra, and email spam identification used in Gmail.

The rest of the paper deals with following sections: Overview of different categories of ML algorithms and its significance in performing decision makingis given in Section I, Section II discusses about the importance of ML in health diagnosis by detecting the disease in early stage and providing customized treatment plan and briefly defined the support of NLP in assisting pharmaceutical industry. Section III presents the literature review of the work carried forwarded by the researchers in this area. Existing challenges of healthcare sector is pointed out in Section IV and Section V involves the applications of ML in addressing the existing challenges. Section VI presents the popular case studies of this sector. Practical implementation of random forest model on diabetic dataset is performed in Section VII. Barriers that are obscuring the complete adoption of ML implementation are mentioned in Section VIII. Lastly, in Section IX, we have concluded our work by stating the promising outcome that will be achieved by intersecting ML with healthcare diagnosis and the limitations that are hindering within the pharma industry.

A. Supervised Learning

The machine learning model is trained with the help of labeled data. The process of gathering raw data such as text, audio, video, or images and the addition of a few appropriate labels in a way that an ML solution can easily identify what data is all about is termed ad data labeling. A training dataset is imported into supervised learning [15]. The training dataset is a large dataset that is used to train ML solutions for predicting the result. The accuracy of any machine learning solution highly depends on the quality and quantity of the training dataset being fed. The dataset contains the input parameters as well as the right outcome [5]. Due to this reason, supervised machine learning is considered to be a learning that is gained under the supervision of an instructor. For example, if an individual wants to know how much duration it will take them to reach their destination, then the labeled data would be whether it is a weekday or weekend, at what time they will depart from the source, and the live weather conditions [16]. If it is a rainy day, then the individual will take longer to reach their destination [6]. The training dataset of the given situation will contain these parameters to predict how much time an individual will take to reach their destination.

B. Unsupervised Learning

In this type of learning, models are not instructed with the help of a training dataset. Machine learning solution themselves looks for the hidden pattern. In other words, the ML model is trained based on an unlabeled dataset [8]. The input parameter is present, but the output is absent inside the dataset. This technique is beneficial for researchers and scientists who are not aware of what they are searching for inside the dataset. Unlabeled data is easily accessible as compared to labeled data. Unknown or hidden insights can be found that are not possible with supervised learning techniques [17]. Labeling of data might cause a manual error, but, in this case, the chances get lowered [7]. Let's assume that an unsupervised learning technique is provided with input that consists of various images of lions and tigers. On this dataset, the algorithm is never trained. The algorithm is not aware of the dataset characteristics. The algorithm will detect similar images and group them all together.

II. Machine Learning in Health Diagnosis

Detection of tumors with accuracy and on time is very much vital to saving human lives in the field of oncology. ML algorithms can identify whether the tumor is benign or malignant in a few seconds. Benign only grows in a specific part of a person's body, and it doesn't turn into cancer. Malignant tumor leads to cancer when the cells grow in multiple numbers and infect the rest of the body parts. World Health Organization (WHO), in their report of the year 2017 for mental illness, mentioned that India has experienced an extreme number of psychological cases as compared to their report in year 1990 [1]. Fig. 1 represents the digital deals that were signed while focusing on enhancing the healthcare sector. Psychological illness has increased since COVID-19 as people have faced personal and economic loss [9]. With biomarkers that help clinicians to identify the disease, ML can identify and analyze who is prone to a specific illness.

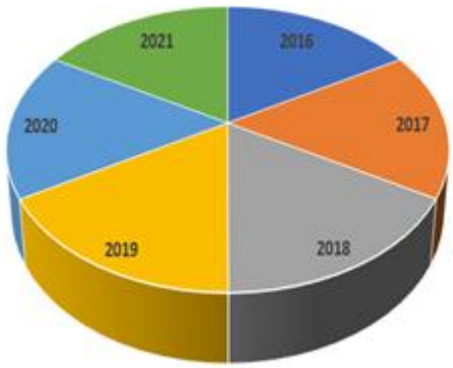

Fig. 1. Digital healthcare deals took place from year 2016 – 2021.

Support Vector Machine (SVM) is categorized under supervised learning to address various diseases. Multiple medical diseases, such as blood pressure and diabetes, can be cured with appropriate computational power and healthcare analytics. Cancer can be diagnosed using SVM [11]. The labeled dataset goes through training and testing datasets, respectively. SVM solution is developed for high accuracy. Based on the mammogram data, detection of breast cancer is done. Factors that are vital to detecting diabetics are blood glucose, age, and body mass index [19]. The SVM model takes these parameters for diagnosing diabetics. Three classes for the output are taken into consideration: diabetic patients, people with a genetic history of diabetes, and non-diabetic people [12]. Regression and classification are used for identifying blood pressure. SVM has flexible implementation and much better performance as compared to the rest of the algorithms. This algorithm is not appropriate when dealing with large amounts of datasets.

A. Motivation

ML has brought down the healthcare cost by replacing manual duties with dedicated technology. Early predication of disease also reduces the chances of complications and preventive initiatives can be taken on time. Worldwide security protocols need to be established to safeguard how ML utilizes data. The quality of health services can be accessed by using image recognition techniques to examine patterns in X-ray reports. New medicine discoveries can be made quickly due to large data gathered from pharmaceutical trials to enhance patients' safety [12]; health tracking can be done actively, which provides pre-emptive recommendations through which future diseases can be avoided. There are various organizations that implement digital healthcare solutions, such as pharmaceuticals, the technological sector, and the government. ML in medical diagnosis is going forward for brining proactive healthcare smart solutions.

B. Research Questions

In this paper, we will look to answer the below questions pertaining to our research.

RQ1.What are the capabilities that can benefit healthcare institutions in terms of accurately providing treatment services [13]? Machine learning facilitates the responsibilities of

medical experts by effectively analysing the health condition and predicting the progression of disease.

RQ2. Is there any real use case of machine learning solutions for delivering quality healthcare? What is the outcome of ML-based health solutions? There are various reputed organizations such as Pizer, Google, IBM, and Tebra that are continuously working on providing health services in different areas and have done extremely well to fulfill the health requirements of end users.

RQ3. Is there any hindrance that is blocking the path of machine learning from getting complete acceptance by stakeholders such as pharma professionals, patients, and third-party organizations? The high performance of machine learning solutions profoundly depends on the quality of the dataset being used. Acknowledging the challenges enables individuals to collect resources that can handle them.

III. Literature Review

Hock Guan Goh et al. [1] presented a machine-learning approach for identifying the growth of diabetes accurately and also predicted complications such as neurological disbalances, heart ailment, and kidney issues that might arise in the future. Various ML techniques, including genetic algorithms, Bayesian networks, and artificial neural networks, are being implemented to treat diabetes. Due to an insufficient volume of historical data for testing and training purposes, there is limited accuracy. The result obtained by applying a large amount of dataset gives a higher performance as compared to those algorithms where only a small quantity of dataset is being applied. Poor output is responsible for high operational charges and also lowers the ML adoption rate. Data fusion is used in the proposed architecture that combines datasets from different sources.

Hamid R. Arabnia et al. [2], in their review paper, discussed the applications of ML for IoT in healthcare. ML can monitor the patient and analyse the health situation depending on the present and previous dataset. The existing health pattern and future complexities of a new disease can be predicted by using a training dataset. The sleeping habit differs from person to person, from a child to a senior citizen. The storage condition of each medication varies in terms of temperature. Annually, there is a huge waste of medicines due to refrigerator failure. IoT sensor technology can be used to track the drug condition and monitory loss can also be eradicated. IoT gathers data from smart devices and makes decisions by using ML algorithms.

Ahmad Shaker Abdalrada et al. [3] worked on cardiac autonomic neuropathy that arises in diabetic patients. Designed an ML-based model to predict the occurrence of this disease in the early stage. Used a dataset containing 200 cardiac autonomic neuropathy test type 2 diabetic patients that are more than 2000. Patients with this disease have increased in past years. Patient records such as gender, age, and health history are stored in the dataset. Parameters such as blood pressure, blood glucose, cholesterol, and body mass index values are also stored. The proposed model has obtained 87% accuracy for the early prediction of this disease.

Ibrahim Mahmood Ibrahim et al. [4] have talked about the key importance of ML in diagnosing disease, and the cooperating sector is working on these techniques for drug discovery. A brief introduction to identifying and predicting diseases with ML is mentioned. Classification algorithms are most commonly applied in the clinical domain that develops training data, and after that, the output is executed on testing data to get precision. The performance of the Support Vector Machine (SVM) degrades when the data load gets increased. Naïve Bayes is suitable for large datasets. K- Nearest Neighbour has complex computation. Decision is used for both classification and regression problems. Random forest consumes lists of time for training. Deep learning can be implemented on different categories of datasets.

Carlo Menon et al. [5] have worked on various ML algorithms by taking multiple bio-signals under consideration. Existing strengths and challenges are discussed that will give insight into future growth in detecting anxiety disorder. The study is done on 102 entities. Reinforcement Learning and SVM are commonly enforced to achieve good output, but the output solely depends on feature selection. Neural Network has given very good output where this method doesn't need any feature selection.

Junhua Yan et al. [6] 2022 worked on a review paper for diagnosing eye diseases such as Glaucoma, and diabetic hypertension that lead to permanent vision loss if not detected in the beginning. ML techniques are being discussed to treat retinal issues that were not implemented in the past. Applications of deep learning models to assist the research in this field are mentioned in depth. A comparative analysis of the background work is given, showing the current and potential scope of ML in eye treatment.

IV. Challenges of Existing Healthcare System

Technology has emerged as a ray of hope to overcome the constraints of the healthcare system. There are qualified healthcare professionals, tech-equipped health devices, and hospitals on one end, and at the same time, the increasing cost of medical facilities exists on the other end [15]. Initially, it is important to accept and acknowledge the challenges of healthcare associated with different parameters. Below, we have discussed the existing healthcare challenges.

A. Cyber Vulnerability

There is inter-connected portable medical equipment that stores patient's medical records. The corporate sector is moving towards digitalization [11]. All India Institute of Medical Sciences, New Delhi, India, faced a ransomware attack at the end of the year 2022 [16]. The attack lasted for 15 days. Records of around 4 crore patients were put at stake, and terabytes of records were encrypted. Cloudsek has disclosed that cyber vulnerabilities in the health industry have witnessed an increase of 95% worldwide in just the beginning four months of the year 2022.

B. Rise Up in Healthcare Costing

Multiple stakeholders, beginning from the manufacturing of equipment and drugs to insurance providers regulate the price of healthcare [17]. Rising costs demoralize people in different aspects, such as undergoing laboratory tests and

regular visits to health practitioners, which affects patient health. Spending a major part of their earning in paying out the medical expenses puts a great burden on people economically [18]. As we know taking prevention is always better than taking cure after a medical situation. The 2019 pandemic has also inflated medical prices globally.

C. Absence of Proper Logistics

Software integrated with AI delivers a huge volume of patient data to pharmaceutical industries. Different data such as patient surveys, transcripts, smart device data, and patient's personal medical data are kept with the health sector. Lack of highly advanced infrastructure leads to mismanagement of data collected from multiple sources [19] [20]. Training of medical staff on the ongoing technology should take place at regular intervals as technology continues to upgrade every time.

V. Emerging Applications of Machine Learning in Healthcare

Healthcare practitioners can deliver effective medical treatment that is customized to particular end users by crushing the vast amount of pharmaceutical data [18]. AI and ML are anticipated to play a vital role in curing chronic diseases. ML based healthcare use cases of year 2020 are presented visually in Fig. 2. ML can help health professionals anticipate the response of patients regarding different medications, which will benefit the organizations in knowing which patients will not face any side effects of the drug. We have below discussed the emerging health applications of ML.

A. Smart Record Management

Powerful medical diagnosis is performed by utilizing huge data to deliver customized end user experience. Enhancing the end-user experience increases the brand value as the company is able to undertake better decision-making [14]. Data entry of patients on online platforms is still arduous and consumes lots of time. Unrevealed hidden parameters are brought into the limelight, which reduces the existing healthcare gap [10]. Electronic Health Record totally depends on the medical records that are inputted inside it. ML can eliminate the challenges of data duplication, billing mistakes, and data loss.

B. Clinical Research

Healthcare companies invest large amounts in finding appropriate individuals who can successfully test new drugs. Maximum percentage of clinical trials results in failure. In the medical field, organizations don't have the option to take risks regarding the outcome of drug trials [7]. If there is a single error in drug development, then it can lead to loss of innocent lives. Appropriate ML algorithms can look for reliable medical testers that can be retained for a longer duration. ML-based clinical trials will enhance the current scenario by examining past trials.

C. Recognizing and Diagnosing Ailments

Earlier, recognizing and curing cancers in the initial stage was very tough. Based on the data entered and the patient's symptoms, ML Can provide diagnostic recommendations. Suitable medications can also be suggested by ML based on the prescription. Side effects of the medications taken can also be predicted [8]. Eko is a healthcare organization in California that is working on AI and ML to fight against lung and heart ailments. SENSORA is an Eko platform for detecting cardiac diseases by bringing machine learning and omnipresent medical equipment together.

VI. Case Studies

Machine learning has provided a remarkable output in terms of processing medical records, disease detection, developing treatment roadmaps, and preventing further complications. Less time consumption, low cost, and efficient management of health data have overall helped doctors to make informed decisions for providing personalized patient care [10]. Below, we have mentioned real-life case studies of renewed organizations that are working on machine learning solutions in the health industry.

A. InnerEye by Microsoft

InnerEye implements ML and computer vision to identify anatomy and tumors separately by harnessing radiological pictures that guide health practitioners in surgical and radiation therapy [17]. The aim of Microsoft is to deliver drugs that are customized according to the requirements of each individual.

B. Datavant Switchboard by Ciox Health

Ciox Health was established in 1976 to work on ML for its product Datavant Switchboard, which allows biopharma groups to quickly retrieve patient information [12]. Strict privacy consent guidelines are adhered to by this organization with regard to patient's health records. The pharmaceutical sector benefits from this platform as industries can develop a customizable control that enables their workers to put forward requests for particular data.

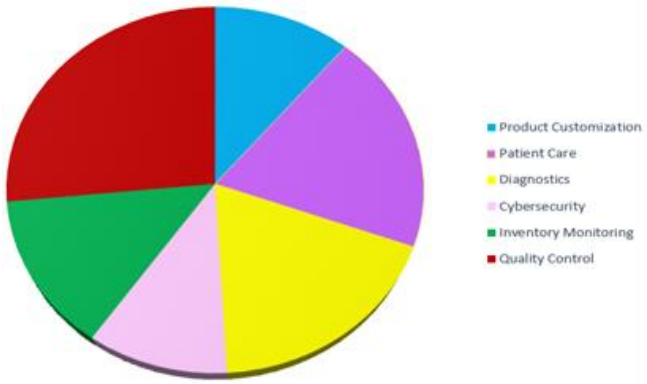

Fig. 2. ML use cases in healthcare in the year 2020.

C. IBM's Watson AI by Pfizer

Natural Language Processing and ML, along with IBM's product Watson AI, are used by Pfizer towards oncological research on how the individual's body can fight back against cancer [13]. Analysis of tons of patients' health information is undertaken by Pfizer to build quick awareness for developing efficient treatment in the field of oncology.

VII. PRACTICAL IMPLEMENTATION

Following are the steps we have implemented to build our random forest classifier with Python.

We have imported the dataset of diabetes prediction from Kaggle. Fig. 3 represents the dataset stored in CSV format, and Fig. 4 shows the successful import of the dataset. There are nine parameters taken into consideration for the dataset, which include pregnancies, glucose, blood pressure, skin thickness, insulin level, body mass index, diabetic pedigree function, age, and outcome. NumPy, pandas, matplotlib, seaborn, sklearn and a few more libraries are used, which is represented in Fig. 5.

	A	B	C	D	E	F	G	H	I
1	Pregnancie	Glucose	BloodPres	SkinThickn	Insulin	BMI	DiabetesPe	Age	Outcome
2	6	148	72	35	0	33.6	0.627	50	1
3	1	85	66	29	0	26.6	0.351	31	0
4	8	183	64	0	0	23.3	0.672	32	1
5	1	89	66	23	94	28.1	0.167	21	0
6	0	137	40	35	168	43.1	2.288	33	1
7	5	116	74	0	0	25.6	0.201	30	0
8	3	78	50	32	88	31	0.248	26	1
9	10	115	0	0	0	35.3	0.134	29	0
10	2	197	70	45	543	30.5	0.158	53	1
11	8	125	96	0	0	0	0.232	54	1
12	4	110	92	0	0	37.6	0.191	30	0

Fig. 3. CSV file.

```
from google.colab import files
uploaded = files.upload()
```

Choose Files Diabetic Dataset.csv

- **Diabetic Dataset.csv**(text/csv) - 23875 bytes, last modified: 3/12/2023 - 100% done

Saving Diabetic Dataset.csv to Diabetic Dataset.csv

Fig. 4. Importing dataset from local device.

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns

sns.set()

from mlxtend.plotting import plot_decision_regions
import missingno as msno
from pandas.plotting import scatter_matrix
from sklearn.preprocessing import StandardScaler
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier

from sklearn.metrics import confusion_matrix
from sklearn import metrics
from sklearn.metrics import classification_report
import warnings
warnings.filterwarnings('ignore')
```

Fig. 5. Importing required libraries.

Fig. 6 deals with reading the dataset that is stored in CSV using the Panda's module. Now we will begin with Exploratory Data Analysis (EDA), which is a technique that deals with analyzing and identifying main parameters in the given dataset using visual approaches such as statical representation and graphs.

```
df_diab = pd.read_csv("Diabetic Dataset.csv")
print(df_diab.info())
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 768 entries, 0 to 767
Data columns (total 9 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
None
```

Fig. 6. Reading the CSV dataset.

Below shows the number of columns and other information regarding dataset is shown in Fig. 7.

```
print(df_diab.describe())
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin
count	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479
std	3.369578	31.972618	19.355807	15.952218	115.244002
min	0.000000	0.000000	0.000000	0.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000
75%	6.000000	140.250000	80.000000	32.000000	127.250000
max	17.000000	199.000000	122.000000	99.000000	846.000000

	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000
mean	31.992578	0.471876	33.240885	0.348958
std	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.078000	21.000000	0.000000
25%	27.300000	0.243750	24.000000	0.000000
50%	32.000000	0.372500	29.000000	0.000000
75%	36.600000	0.626250	41.000000	1.000000
max	67.100000	2.420000	81.000000	1.000000

Fig. 7. Describing the dataset.

While collecting data from multiple sources, many times missing values lead to low performance of the ML model. Missing values or null values are very common. Checking for null values in the dataset by using the isnull method, which is predefined. To see the first 8 rows with null values, the head function is used (Fig. 8).

```
df_diab.isnull().head(8)
```

1 to 8 of 8 entries Filter

index	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	false	false	false	false	false	false	false	false	false
1	false	false	false	false	false	false	false	false	false
2	false	false	false	false	false	false	false	false	false
3	false	false	false	false	false	false	false	false	false
4	false	false	false	false	false	false	false	false	false
5	false	false	false	false	false	false	false	false	false
6	false	false	false	false	false	false	false	false	false
7	false	false	false	false	false	false	false	false	false

Fig. 8. Checking for null values.

As dataset contains null values, so will take the sum of those null values using sum function. In Fig. 9, all the null values are denoted with 0.

```
df_diab.isnull().sum()
```

```
Pregnancies                 0
Glucose                     0
BloodPressure               0
SkinThickness               0
Insulin                     0
BMI                         0
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Fig. 9. Number of null values.

Fig. 10 shows the replaced null values that were represented 0 with NaN (Not a Number).

```
[40] df_diab_copy = df_diab.copy(deep = True)
     df_diab_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']] = df_diab_copy[['Glucose','BloodPressure','SkinThickness','Insulin','BMI']].replace(0

[41] print(df_diab_copy.isnull().sum())

Pregnancies                 0
Glucose                     5
BloodPressure              35
SkinThickness             227
Insulin                   374
BMI                        11
DiabetesPedigreeFunction    0
Age                         0
Outcome                     0
dtype: int64
```

Fig. 10. Replacing the null values.

Data distribution is done in Fig. 11 before removal of null values.

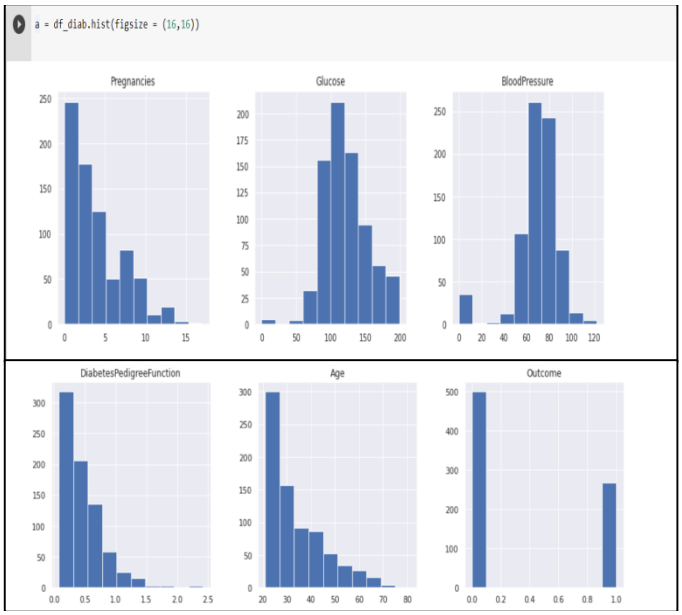

Fig. 11. Plotting data distribution with null values.

Fig. 12 deals with inserting the mean values of the columns where the missing values were identified in previous steps.

```
df_diab_copy['Glucose'].fillna(df_diab_copy['Glucose'].mean(), inplace = True)
df_diab_copy['BloodPressure'].fillna(df_diab_copy['BloodPressure'].mean(), inplace = True)
df_diab_copy['SkinThickness'].fillna(df_diab_copy['SkinThickness'].median(), inplace = True)
df_diab_copy['Insulin'].fillna(df_diab_copy['Insulin'].median(), inplace = True)
df_diab_copy['BMI'].fillna(df_diab_copy['BMI'].median(), inplace = True)
```

Fig. 12. Setting the mean values.

Plot distribution after removal of null values is shown in Fig. 13.

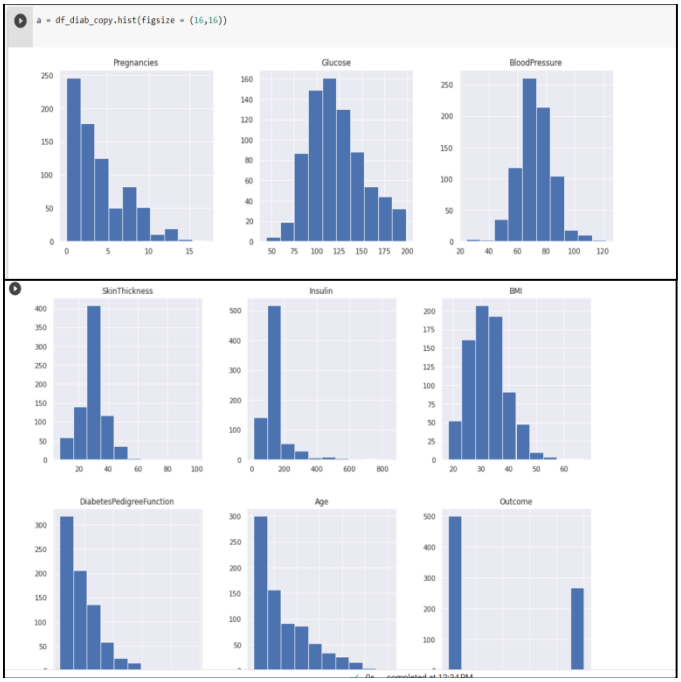

Fig. 13. Distributing after removing null values.

Below bar graph shows that no null values exist in the dataset (Fig. 14).

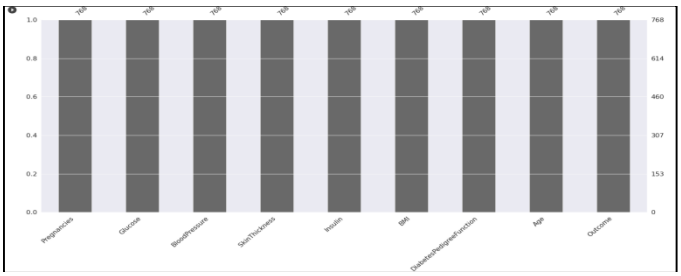

Fig. 14. Counting null values analysis.

The total number of diabetic patients are half of the rest of non-diabetic patients (see Fig. 15 and 16).

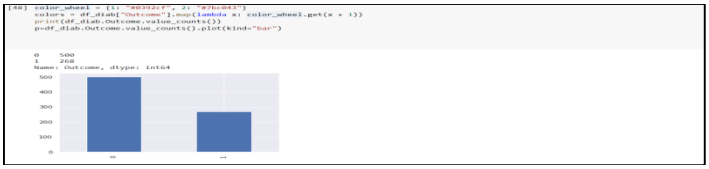

Fig. 15. Imbalance data.

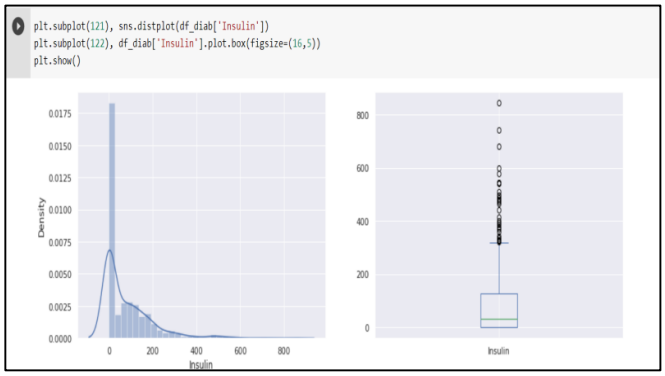

Fig. 16. Boxplot.

Correlation of all the parameters before data cleaning is shown in Fig. 17 and 18.

```
plt.figure(figsize=(12,10))
p = sns.heatmap(df_diab.corr(), annot=True,cmap ='RdYlGn')
```

Fig. 17. Correlation among all features.

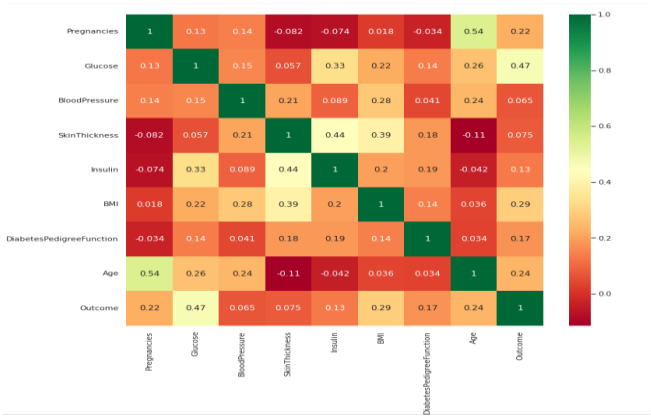

Fig. 18. Correlation.

[52] df_diab_copy.head()

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148.0	72.0	35.0	125.0	33.6	0.627	50	1
1	1	85.0	66.0	29.0	125.0	26.6	0.351	31	0
2	8	183.0	64.0	29.0	125.0	23.3	0.672	32	1
3	1	89.0	66.0	23.0	94.0	28.1	0.167	21	0
4	0	137.0	40.0	35.0	168.0	43.1	2.288	33	1

Fig. 19. Before scaling.

```
[53] sc_X = StandardScaler()
     X =  pd.DataFrame(sc_X.fit_transform(df_diab_copy.drop(["Outcome"],axis = 1),), columns=['Pregnancies',
     'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age'])
     X.head()
```

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
0	0.639947	0.865108	-0.033518	0.670643	-0.181541	0.166619	0.468492	1.425995
1	-0.844885	-1.206162	-0.529859	-0.012301	-0.181541	-0.852200	-0.365061	-0.190672
2	1.233880	2.015813	-0.695306	-0.012301	-0.181541	-1.332500	0.604397	-0.105584
3	-0.844885	-1.074652	-0.529859	-0.695245	-0.540642	-0.633881	-0.920763	-1.041549
4	-1.141852	0.503458	-2.680669	0.670643	0.316566	1.549303	5.484909	-0.020496

Fig. 20. After scaling.

The dataset is being splatted into training and testing datasets (Fig. 21 and 22).

```
[54] X = df_diab.drop('Outcome', axis=1)
     y = df_diab['Outcome']
```

Fig. 21. Building model.

```
[57] from sklearn.model_selection import train_test_split

     X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.33,
                                                         random_state=7)
```

Fig. 22. Dataset splitting.

Fig. 25 examines the accuracy of the random forest model on the training dataset.

```
[58] from sklearn.ensemble import RandomForestClassifier

     rfc = RandomForestClassifier(n_estimators=200)
     rfc.fit(X_train, y_train)

     RandomForestClassifier
     RandomForestClassifier(n_estimators=200)
```

Fig. 23. Random forest model building.

```
from sklearn import metrics
```

Fig. 24. Importing necessary metrics.

```
[60] predictions = rfc.predict(X_test)
     print("Accuracy_Score =", format(metrics.accuracy_score(y_test, predictions)))

     Accuracy_Score = 0.7637795275590551
```

Fig. 25. Accuracy.

We have implemented a random forest model for predicting diabetics (Fig. 23 and 24). Various parameters including age, insulin level, BMI, blood pressure are taken into consideration. Fig. 26 shows the flowchart of the random forest-based prediction model. Data distribution of these parameters is represented graphically. Null values are eliminated from the dataset after setting up the mean values. This shows that the total strength of diabetic patients is half of the total number of non-diabetic patients. After scaling, the dataset is divided into testing and training dataset where all the features have a correlation coefficient of 1, represents that all the features present in the dataset are positively correlated with rest of the features (see Fig. 19 and 20). Accuracy of the model is 0.76377952755, which represents a high accuracy.

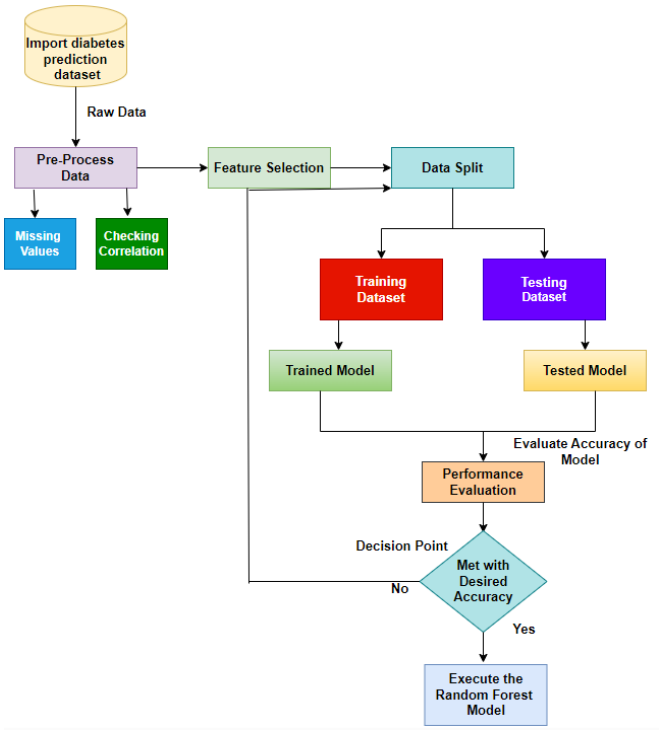

Fig. 26. Flowchart of prediction model.

VIII. Barriers in Machine Learning Implementation

Primary care mistake has an edge of error. Lots of patients' lives can be put at stake just because of a single error. Radiological scans based on machine learning might fail to scan a tumor [4]. More questions could arise if a user consumes the wrong drug recommended by the healthcare model [10]. Table I presents the list of companies that are working on healthcare sector. Machine learning can manage tasks quickly that were earlier handled by humans manually.

A. Privacy Invasion

Privacy remains a topic of concern regarding patient data. Even after applying different measures, the presence of malicious attackers that are trying hard to breach the security of smart systems. Personal data of patients carries very sensitive identification details, including bill payment information [6]. Machine learning has the ability to predict patient information even when the model is not fed with any data. The University of Washington faced a security incident in the year 2022 when fraudulently gained access to the university network system.

B. Bias Nature

Discrimination in the machine learning model occurs when data is imported from a particular source, and the output that is obtained can only be beneficial for that particular source [17]. Suppose that data imported from an academic health science center is fed to the ML system. That system might not give fruitful results to those population that belongs to the rest of the areas except for the academic medical center [18]. Similarly, while using speech recognition in recording notes, the model might show low efficiency when the user belongs to a different race.

C. Faulty Diagnosis

We are already aware that the quality of the machine learning model entirely depends on the dataset being provided. There is the potential risk of a negative diagnosis [8]. The dataset being imported might not contain a sufficient amount of information from different socio-economic backgrounds. Medical professionals might be held accountable if the ML model makes a wrong decision because the doctor was the one who used the smart model to make health decisions.

TABLE I. Companies Utilizing Technology in Healthcare

S.No.	Healthcare Companies	Area of Treatment	Technology Leveraged
1	Google DeepMind	Breast cancer diagnosis	Machine Learning and Image Processing
2	IBM Watson Health	Breast cancer treatment and faster drug delivery	Cognitive Computing
3	CloudMedX hEALTH	Heart failure and liver cancer	Natural Language Processing and Deep Learning
4	Oncora Medical	Radiation therapy for cancer treatment	Machine Learning and Natural Language Processing
5	Babylon Health	Primary Healthcare	Deep Learning
6	Corti	Cardiac disease	Artificial Neural Network
7	Butterfly Network	Ultrasound and MRI examination	Artificial Intelligence and Cloud Computing

IX. Conclusion and Future Scope

The objective of this study was to analyze different ML techniques in revolutionizing medical diagnosis. To achieve the same, we have accessed various literature work starting from year 2018-2023. We have recognized four major databases: Springer Link, IEEE, IGI Global and De Gruyter. We have examined the key benefits of ML in eliminating the existing healthcare challenges. The studies done so far has shown significant improvement in their results, Our study have shown that ML not only brings down the entire treatment cost but along with this recognizes the hidden pattern that indicate disease in initial stage itself. Machine learning has been disruptive technology, from predicting diseases in the early stages by examining radiological images to moving towards a fast, efficient, and smart healthcare system that can become the savior of tons of human lives. We have discussed about the most popular case studies of pharma industry and how those solutions are assisting individuals in tackling medical conditions. Implemented random forest model for diabetic's prediction which showed an accuracy of 76377952755. This study would give researchers a primary knowledge to carry forward their work. We have not considered the work that is presented in any other language. So in the upcoming years we can consider the resources that have been neglected as those can also provide valuable insights.

Accurate medical diagnosis and personalized health treatment have excessively refined medical research. Its capability to quickly analyses large quantities of clinical records assists medical practitioners in recognizing disease in the early stage. Although there are potential challenges, it is

clearly visible that machine learning will lay a foundation for enhancing the worldwide health ecosystem. Together with machine learning, people are conscious of the importance of a healthy lifestyle. By mitigating treatment costs and high-quality patient care, machine learning has impressively transformed the healthcare system and human lives as well.

REFERENCES

[1] Nadeem, Muhammad Waqas, Hock Guan Goh, Vasaki Ponnusamy, Ivan Andonovic, Muhammad Adnan Khan, and Muzammil Hussain. "A fusion-based machine learning approach for the prediction of the onset of diabetes." In *Healthcare*, vol. 9, no. 10, p. 1393. MDPI, 2021.

[2] Mohammadi, Farid Ghareh, Farzan Shenavarmasouleh, and Hamid R. Arabnia. "Applications of machine learning in healthcare and internet of things (IOT): a comprehensive review." *arXiv preprint arXiv:2202.02868* (2022).

[3] Abdalrada, Ahmad Shaker, Jemal Abawajy, Tahsien Al-Quraishi, and Sheikh Mohammed Shariful Islam. "Prediction of cardiac autonomic neuropathy using a machine learning model in patients with diabetes." *Therapeutic Advances in Endocrinology and Metabolism* 13 (2022): 20420188221086693.

[4] Ibrahim, Ibrahim, and Adnan Abdulazeez. "The role of machine learning algorithms for diagnosing diseases." *Journal of Applied Science and Technology Trends* 2, no. 01 (2021): 10-19.

[5] Ancillon, Lou, Mohamed Elgendi, and Carlo Menon. "Machine learning for anxiety detection using biosignals: a review." *Diagnostics* 12, no. 8 (2022): 1794.

[6] Abbas, Qaisar, Imran Qureshi, Junhua Yan, and Kashif Shaheed. "Machine learning methods for diagnosis of eye-related diseases: a systematic review study based on ophthalmic imaging modalities." *Archives of Computational Methods in Engineering* 29, no. 6 (2022): 3861-3918.

[7] Ahsan, Md Manjurul, and Zahed Siddique. "Machine learning-based heart disease diagnosis: A systematic literature review." *Artificial Intelligence in Medicine* 128 (2022): 102289.

[8] Chittora, Pankaj, Sandeep Chaurasia, Prasun Chakrabarti, Gaurav Kumawat, Tulika Chakrabarti, Zbigniew Leonowicz, Michał Jasiński et al. "Prediction of chronic kidney disease-a machine learning perspective." *IEEE access* 9 (2021): 17312-17334.

[9] Marwan, Mbarek, Ali Kartit, and Hassan Ouahmane. "Security enhancement in healthcare cloud using machine learning." *Procedia Computer Science* 127 (2018): 388-397.

[10] S. Bharany, K. Kaur, S. E. M. Eltaher, A. O. Ibrahim, S. Sharma, and M. M. M. A. Elsalam, "A Comparative Study of Cloud Data Portability Frameworks for Analyzing Object to NoSQL Database Mapping from ONDM's Perspective," International Journal of Advanced Computer Science and Applications, vol. 14, no. 10. The Science and Information Organization, 2023. doi: 10.14569/ijacsa.2023.0141086.

[11] Gupta, Medini, Sarvesh Tanwar, Ajay Rana, and Himdweep Walia. "Smart healthcare monitoring system using wireless body area network." In *2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pp. 1-5. IEEE, 2021.

[12] A. Sundas, S. Badotra, S. Bharany, A. Almogren, E. M. Tag-ElDin, and A. U. Rehman, "HealthGuard: An Intelligent Healthcare System Security Framework Based on Machine Learning," Sustainability, vol. 14, no. 19. MDPI AG, p. 11934, Sep. 22, 2022. doi: 10.3390/su141911934.

[13] Gupta, Medini, Sarvesh Tanwar, Sumit Badotra, and Ajay Rana. "A systematic review on blockchain in transforming the healthcare sector." *Transformations Through Blockchain Technology: The New Digital Revolution* (2022): 181-200.

[14] V. Sapra et al., "Integrated approach using deep neural network and CBR for detecting severity of coronary artery disease," Alexandria Engineering Journal, vol. 68. Elsevier BV, pp. 709–720, Apr. 2023. doi: 10.1016/j.aej.2023.01.029.

[15] Tanwar, Sarvesh, Neelam Gupta, Celestine Iwendi, Karan Kumar, and Mamdouh Alenezi. "Next generation IoT and blockchain integration." *Journal of Sensors* 2022 (2022).

[16] Badotra, Sumit, Sarvesh Tanwar, Ajay Rana, Nidhi Sindhwani, and Ramani Kannan, eds. *Handbook of augmented and virtual reality*. De Gruyter, 2023.

[17] K. Kaushik et al., "Multinomial Naive Bayesian Classifier Framework for Systematic Analysis of Smart IoT Devices," Sensors, vol. 22, no. 19. MDPI AG, p. 7318, Sep. 27, 2022. doi: 10.3390/s22197318.

[18] Tanwar, Sarvesh, Sumit Badotra, Medini Gupta, and Ajay Rana. "Efficient and secure multiple digital signature to prevent forgery based on ECC." *International Journal of Applied Science and Engineering* 18, no. 5 (2021): 1-7.

[19] K. Kaushik et al., "A Machine Learning-Based Framework for the Prediction of Cervical Cancer Risk in Women," Sustainability, vol. 14, no. 19. MDPI AG, p. 11947, Sep. 22, 2022. doi: 10.3390/su141911947.

[20] Hickman, Sarah E., Ramona Woitek, Elizabeth Phuong Vi Le, Yu Ri Im, Carina Mouritsen Luxhøj, Angelica I. Aviles-Rivero, Gabrielle C. Baxter, James W. MacKay, and Fiona J. Gilbert. "Machine learning for workflow applications in screening mammography: systematic review and meta-analysis." *Radiology* 302, no. 1 (2022): 88-104.