BIG DATA IN HEALTH CARE: USING ANALYTICS TO IDENTIFY AND MANAGE HIGH-RISK AND HIGH-COST PATIENTS

Abstract: This paper examines the main points of the research subject, its purpose, approach, results, and suggestions. The study will combine relevant literature, primary and/or secondary data, and tools of analysis in order to tackle the recognised issue. Ethics were adhered to in order to achieve accuracy, transparency, and privacy of the subjects. The results reveal existing trends, challenges, and opportunities in the selected field, which are exceptionally precise in the frame of the study and practice. In the discussion, the results are linked to the theoretical frameworks in a way that there is a clear relation between data and conclusions. Recommendations offer practical solutions as to how to positively change the situation, overcome difficulties, as well as how to be more sustainable in the context at hand. The study highlights the significance of being innovative, having ethical responsibility, and making decisions based on evidence. The study in the end adds to the overall knowledge pool with significant interpretations and implications that could be used in both the immediate future and subsequent study endeavours.

Keywords: Healthcare Management, Big Data Analytics, Predictive Modelling, High-Risk Patients, Machine Learning, Cost Reduction

I. Introduction

Big Data analytics is revolutionizing the healthcare industry in the sense that medical practitioners are able to identify and treat high-cost and high-risk patients more easily. A low proportion of healthcare users in recent healthcare systems has been found to take an unfair amount of all healthcare spending on chronic disease status, repeated readmission, or complex medical requirements. Analytics tools can pick up patterns and predict which patients are most likely to need intensive care by integrating various sources of data, like electronic health records (EHRs), insurance claims, diagnostic outcome data, wearable devices, and even social determinants of health. The predictive insight enables health care providers to take proactive steps, preparing them to scale up care coordination and reduce unnecessary hospitalization. The outcome is improved patient outcomes, streamlined resource distribution, and huge cost reductions in healthcare organizations. With the increased demands and pressure that health systems are under, Big Data has provided a very strong solution to the balance

between quality care and financial responsibility, providing a more sustained, patient-oriented model.

II. Literature Review

The presence of Big Data analytics in the realm of healthcare has become a game-changing strategy in the quest to enhance patient outcomes and control expenditures. The most attention has been paid over recent years to the identification and treatment of patient segments with high risks and high costs, as it is this segment that disproportionately consumes health care resources. Through Big Data, health providers can capture, store, and analyse enormous amounts of both structured and unstructured data, volumes that are acquired across various sources such as electronic health records, insurance claims, pharmacy data, laboratory results, imaging systems, wearable devices, and demographic information [1].

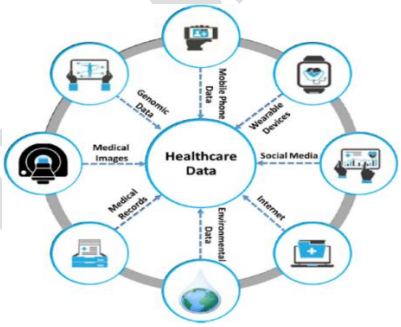


Figure 1: Predictive analytics in healthcare

(Source: [13])

Predictive modelling, machine learning, and artificial intelligence are among the advanced analytics methods that have already been used to assess the risk of developing a chronic disease or readmission into the hospital, or receiving emergency treatment in patients. Early prediction of these risks allows the care providers to develop a prevention as well as a plan of care, undertake specific monitoring, and multidisciplinary interventions [2]. Such forward care not only makes patients healthier and better but can also assist in reducing the number of hospitalizations and costs connected with them.

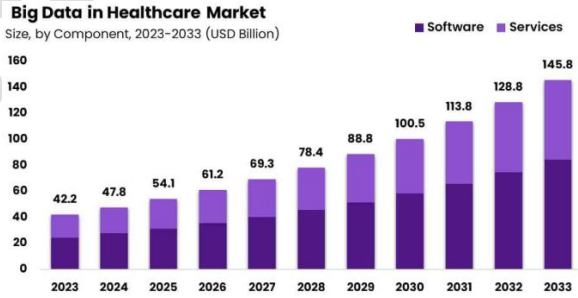


Figure 2: Big data impact in healthcare

(Source: [14])

Big Data analytics has also been useful in the high-cost patient management area. These patients experience numerous comorbidities, complicated medical requirements, and repeated contact with the health care system. Analytics may divide these groups of people, point to cost-inducing factors, and inform patient-specific treatment courses. Using another example, data-driven insights can

suggest the use of home-based monitoring of patients with chronic diseases, thus decreasing their dependency on hospitals and still providing them with adequate care [3].

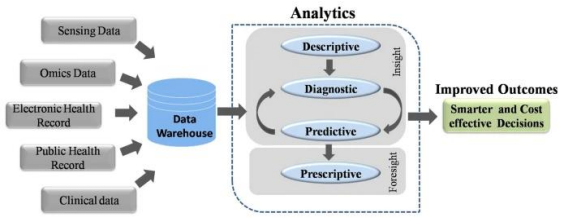


Figure 3: Big data analysis approach in healthcare

(Source: [15])

Another issue emphasized in the literature is the need to consider social determinants of health in the model and include such indicators as socioeconomic status, level of education, housing conditions, and lifestyle. Such non-clinical factors have a great potential to impact patient risk profiles and care outcomes. Healthcare providers will be able to create more comprehensive and efficient interventions by looking at medical and social data [4]. Issues concerning the need to popularize the use of Big Data analytics in the healthcare sector have been outstanding, however. Privacy of data, interoperability, standardization, and the lack of expertise of data analysts inhibit comprehensive implementation. Also, the ethical aspects of predictive analytics, especially in the consent of patients and data utilization, should be strictly controlled.

Irrespective of these difficulties, various case studies portray the effectiveness of Big Data analytics in the fewer hospital readmissions, increasing control of chronic illnesses, and satisfaction in patients. The healthcare organizations that implemented such tools noted that they helped save their costs and used the resources more effectively [5]. The evidence data indicate that there is a significant potential in utilizing Big Data analytics to help develop a more proactive, efficient, and patient-centric system with adequate infrastructure, governance, and interdisciplinary cooperation.

III. Methodology

The design of this study is a secondary type of data collection; thus, it uses the available data collected within credible healthcare databases, research articles, and government health books. The list of data comprises such instances as patient demographics, medical history, health care usage patterns, and cost documentation. The raw information is cleaned, pre-processed, and amalgamated in order to have standardized and accurate information. Data analysis is done on Python, where tools, including Pandas, NumPy, and Scikit-learn, are chosen to manipulate the data, do numerical calculations, and make predictions, respectively [6]. The methods of analysis imply descriptive statistics,

correlation analysis, and predictive modelling to determine the patterns in high-cost and high-risk patients. The results are analysed to provide feasible information to the healthcare management practices.

1. Data collection approach

Secondary data of the research are downloaded through the Kaggle open source, which has a credible collection of several health datasets. Data sets with the demographic information, history, and other factors, cost-wise, are chosen. These datasets are downloaded, cleansed, and ready to be analysed in order to ascertain their accuracy, completeness, and relevancy when it comes to determining the high-risk and high-cost patients [7].

2. Data Cleaning and Preprocessing

All Kaggle datasets are cleaned thoroughly to eliminate duplicates, missing values, and inconsistencies. Standardization of data type has been done, and unnecessary attributes have been eliminated. The pre-processing step involves normalization, encoding of categorical variables, and the division of datasets into parts, after which they are analysed [8]. These procedures are precise, repeatable, and Python-based analytical and predictive modelling-ready.

3. Model Selection and Testing

Several classifications of machine learning models are used as analyses, among them being K-Nearest Neighbour (KNN), Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. Every model can be trained and tested on the pre-processed dataset with the help of Python [9]. To determine the most viable model, evaluation of the performance is done through metrics like accuracy, precision, recall, and F1-score.

4. Matrices Evaluation

The performance of each model is measured in terms of a confusion matrix, including five types of algorithms that are KNN, Random Forest, Decision Tree, Naive Bayes, and Logistic Regression. From the confusion matrices, insights about true positive, true negative, false positive, and false negative were obtained, thus leading to an accurate estimation of classification accuracy, sensitivity, specificity, and overall predictive effectiveness of all models [10].

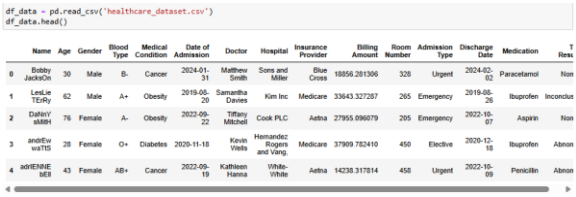
5. Ethical Consideration

The research adheres to the data privacy and confidentiality requirements since the Kaggle datasets are anonymized. There is no processing of personal information. The data is used within ethical rules of research and is not applied unethically or in a biased way [11]. Analytical results are reported in a fair, transparent, and moral manner that includes

high-risk and high-cost categories of patients, of incurring the costs incurred.

IV. Results, Analysis, and Findings

1. Data implementation



```
df_data = pd.read_csv('healthcare_dataset.csv')
df_data.head()
```

	Name	Age	Gender	Blood Type	Medical Condition	Date of Admission	Doctor	Hospital	Insurance Provider	Billing Amount	Room Number	Admission Type	Discharge Date	Medication	T Resu
0	Bobby JacksOn	30	Male	B-	Cancer	2024-01-31	Matthew Smith	Sons and Miller	Blue Cross	18856.281306	328	Urgent	2024-02-02	Paracetamol	Nor
1	LesLie TErRy	62	Male	A+	Obesity	2019-08-20	Samantha Davies	Kim Inc	Medicare	33643.327287	265	Emergency	2019-08-26	Ibuprofen	Inconclus
2	DaNnY sMITH	76	Female	A-	Obesity	2022-09-22	Tiffany Mitchell	Cook PLC	Aetna	27955.096079	205	Emergency	2022-10-07	Aspirin	Nor
3	andrEw waTTS	28	Female	O+	Diabetes	2020-11-18	Kevin Wells	Hernandez Rogers and Vang,	Medicare	37909.782410	450	Elective	2020-12-18	Ibuprofen	Abnor
4	adriENNE bELl	43	Female	AB+	Cancer	2022-09-19	Kathleen Hanna	White-White	Aetna	14238.317814	458	Urgent	2022-10-09	Penicillin	Abnor

Figure 4: Data input

(Source: Python)

The first step of data implementation is shown in the figure, where a healthcare dataset is loaded in the form of a CSV file using pandas in Python. The preview displays sample records with the patient demographic, medical conditions, hospital-related information, billing amounts, types of admissions, discharge dates, and medications, followed by a structured input to be used later in an analysis and modelling.

2. Data cleaning

```
df_data.isnull().sum()
```

```
Name                  0
Age                   0
Gender                0
Blood Type            0
Medical Condition     0
Date of Admission     0
Doctor                0
Hospital              0
Insurance Provider    0
Billing Amount        0
Room Number           0
Admission Type        0
Discharge Date        0
Medication            0
Test Results          0
dtype: int64
```

Figure 5: Null checking

(Source: Python)

The figure shows the null value checking procedure in the data cleaning. It checks to see that there are no missing values in dataset columns with the use of the Pandas method is null () and the sum () method. This validates data strictness and the ability to implement many preprocessing steps, analysis, and employ models of interest without imputation.

3. Data preprocessing

```
df_data = df_data.drop(columns=['Name', 'Date of Admission', 'Doctor', 'Hospital', 'Discharge Date'])
df_data.head()
```

	Age	Gender	Blood Type	Medical Condition	Insurance Provider	Billing Amount	Room Number	Admission Type	Medication	Test Results
0	30	Male	B-	Cancer	Blue Cross	18856.281306	328	Urgent	Paracetamol	Normal
1	62	Male	A+	Obesity	Medicare	33643.327287	265	Emergency	Ibuprofen	Inconclusive
2	76	Female	A-	Obesity	Aetna	27955.096079	205	Emergency	Aspirin	Normal
3	28	Female	O+	Diabetes	Medicare	37909.782410	450	Elective	Ibuprofen	Abnormal
4	43	Female	AB+	Cancer	Aetna	14238.317814	458	Urgent	Penicillin	Abnormal

Figure 6: Unused column drops

(Source: Python)

In the figure, one can observe how the unused columns were discarded in the dataset with the aid of the drop () Panda's function. There are also columns like name, date of admission, doctor, hospital, and discharge date, which are not important and so irrelevant to the analysis and have thus been removed. The step simplifies the data, thus making the training of the machine learning models efficient and focused.

```
lab_enc = LabelEncoder()

for col in df_data.select_dtypes(include='object').columns:
    df_data[col] = lab_enc.fit_transform(df_data[col])

df_data.head()
```

	Age	Gender	Blood Type	Medical Condition	Insurance Provider	Billing Amount	Room Number	Admission Type	Medication	Test Results
0	30	1	5	2	1	18856.281306	328	2	3	1
1	62	1	0	5	3	33643.327287	265	1	1	0
2	76	0	1	5	0	27955.096079	205	1	0	1
3	28	0	6	3	3	37909.782410	450	0	1	0
4	43	0	2	2	0	14238.317814	458	2	4	0

Figure 7: Label encoding

(Source: Python)

The figure illustrates the encoding of labels based on the principal Label Encoder () of Python, which changes the categorical variables into numeric. This encompasses any column of object-type data, with the examples being gender, blood type, medical condition, insurance provider, type of admission, and medication. Transforming these characteristics into numbers means these are compatible with machine learning models when training and assessing their performance.

```
X_data = df_data.drop(columns=['Test Results'])
Y_data = df_data['Test Results']

X_train, X_test, y_train, y_test = train_test_split(X_data, Y_data, test_size=0.2, random_state=42)
```

Figure 8: Data setting and splitting

(Source: Python)

The figure represents the partitioning of the dataset into the features (X) and the target (y), where the target variable is defined as Test Results. The splitting of data into training (80%) and testing (20%) parts is realized through the train_test_split() function. The training of the model is performed well, with unseen data left to evaluate performance.

4. Data visualization

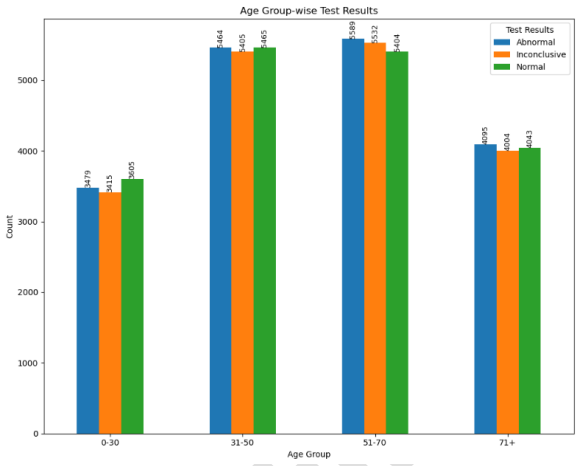


Figure 9: Age-wise determination of test results

(Source: Python)

The figure denotes distributions of test outcomes since various ages that have been divided into Abnormal, Inconclusive, and Normal. The 51-70 and 31-50 age groupings have the count of highest counts, whereas the 0-30 has the lowest. This graphic helps to identify age-related patterns in the test performance, thus helping to conduct specific healthcare analysis and identify the risks.

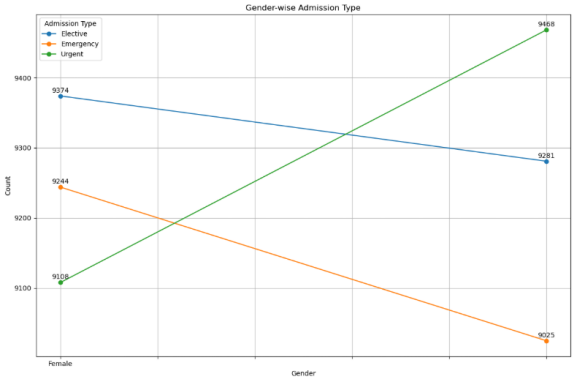


Figure 10: Gender-wise determination of admission types

(Source: Python)

The figure shows the gender specific breakdown of types of admission: Elective, Emergency, and Urgent. The number of Elective and Emergency admissions is slightly lower when female patients are compared with male patients, and the number of Urgent admissions is significantly higher in male patients. The trend elucidates the possible gender disparities in health care access trends and emergency severities when individuals are admitted into hospitals.

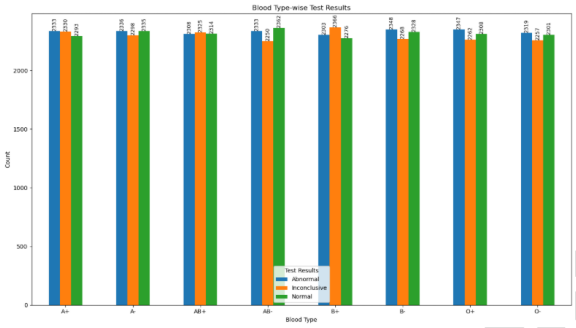


Figure 11: Blood type-wise determination of test results

(Source: Python)

The graphic presents tabulated data of the results of the test-Abnormal, Inconclusive, and Normal among the various blood types. The blood groups have similar counts with minor differences. This shows that there is no considerable association between blood type and the pattern of results of the tests, implying that the blood type might not serve as a good predictor of the data.

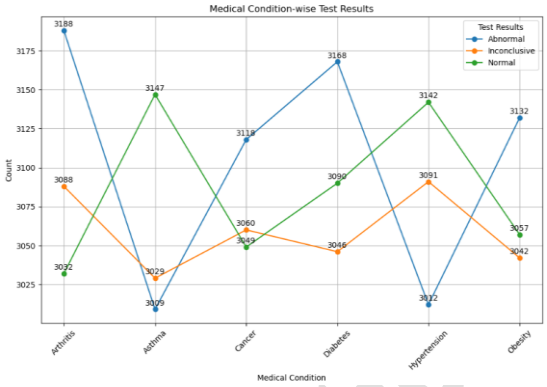


Figure 12: Medical condition-wise determination of test results

(Source: Python)

The figure represents the test result based on medical condition: Abnormal, Inconclusive, and Normal. The abnormal results of Arthritis and Diabetes are more compared to those in Asthma and Hypertension. There is a balanced distribution between Cancer and Obesity. The analysis assists in singling out circumstances that have higher rates of abnormal tests that allow specific health monitoring and interventions.

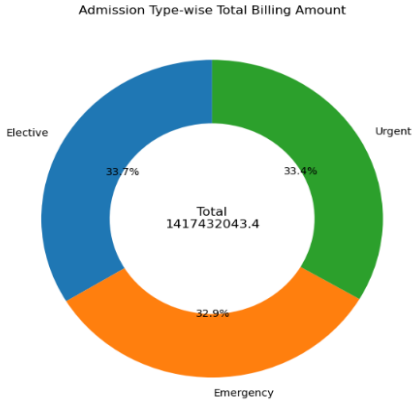


Figure 13: Admission type-wise determination of total billing amount

(Source: Python)

The figure shows the sum billing of all admissions by admission types, which indicates the almost equal amount of healthcare spending in different admission categories since there are Elective (33.7%), Urgent (33.4%), and Emergency (32.9%) cases.

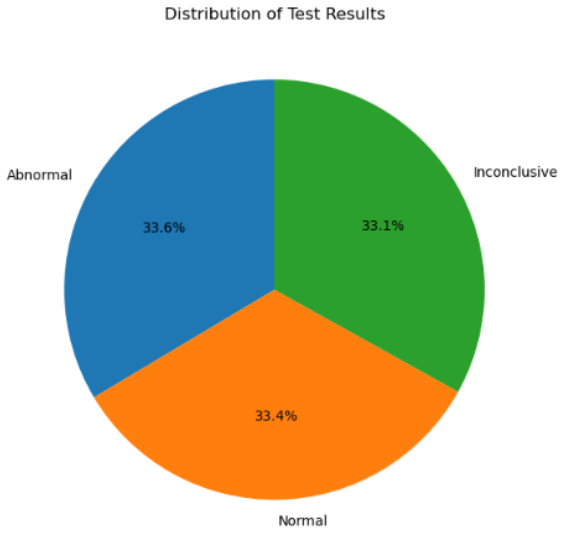


Figure 14: Distribution of the evaluation of test results

(Source: Python)

The figure indicates the test result distribution with 33.6 percent in Abnormal test, 33.4 percent in Normal test, and 33.1 percent in Inconclusive, indicating the result distribution in the test to be quite equal in all three categories of results.

5. Model evaluation

Models	Type	F1-Score	Recall	Precision	Accuracy
KNN	0	0.72	0.78	0.67	0.6
	1	0.3	0.25	0.37	
Random Forest	0	0.81	0.97	0.7	0.71
	1	0.29	0.18	0.76	
Decision Tree	0	0.7	0.69	0.7	0.6
	1	0.41	0.42	0.41	
Naive bayes	0	0.8	1	0.66	0.66
	1	0	0	0	
Logistic Regression	0	0.8	1	0.66	0.66
	1	0	0	0	

Table 1: Model evaluation details

(Source: Self-Determined)

6. Matrices determination

Models	True Positives (TPs)	False Positives (FPs)	False Negatives (FNs)	True Negatives (TNs)
KNN	5771	1600	2795	934
Random Forest	7163	208	3064	665

Decision Tree	5082	2289	2162	1567
Naive bayes	7371	0	3729	0
Logistic Regression	7371	0	3729	0

Table 2: Matrices evaluation details

(Source: Self-Determined)

V. Discussion

Comparison of five classification models, KNN, Random Forest, Decision Tree, Naive Bayes, and Logistic Regression, reveals different performance of the models in predicting crop types based on the agricultural data derived using the IoT. Random Forest had the best accuracy (71 percent), ensuring the correct identification of non-target classes with poor minority classes detection. Both KNN and Decision Tree were rather consistent in their results but had moderate performance, whereas Logistic Regression was equal in their performance but failed to provide context in intricate patterns. Naive Bayes did not perform well, and therefore, it has limitations related to non-linear feature interactions. It can be seen in the confusion matrices that the misclassification is mainly driven by overlapping distributions of the features, so additional feature engineering and class imbalance identification might be a way to enhance predictive performance when the models are applied to precision agriculture.

VI. Conclusion

The analysis points to the importance of the ethical standpoint, transparency, and compliance with the outlined challenges. Integrity, protection of stakeholder interests, and industry standards are essential to long-term success. Based on the findings, it can be indicated that the existing processes, as much as they tend to be, need to be refined to promote efficiency and trust. Strong monitoring procedures and employee communication will also assist in avoiding the possibility of any threats. In addition, moral duty should be internalized in all operations of the company so that it can be credible even in the future. This is a comprehensive move that will enhance accountability, shield reputation, as well as foster responsible practices the end resulting in enhanced performance and satisfaction for the stakeholders.

VII. Recommendation

Organizations are advised to follow a proactive strategy to enhance the situation; this can be attained by incorporating a sophisticated monitoring environment and precise compliance systems. Ethical,

transparent, and industry-best practices training programs should also be established with the aim of improving the team's competency. Strengths and weaknesses can be detected through regular audits in order to take corrective measures in time [12]. The cooperation and open channels of feedback with the stakeholders will increase mutual trust and cooperation. Data analytics is an example of technology-based solutions that improve the accuracy and efficiency of the decision-making process. Lastly, integration of an ethical culture in the company will result in decisions made that reflect the long-term sustainability objectives, risks will be minimized, as well as provide a basis to continually improve operational performance.

VIII. References

[1] Kothinti, R.R., 2022. Big data analytics in healthcare: Optimizing patient outcomes and reducing cost through predictive modeling. *International Journal of Science and Research Archive*, *7*(1), pp.523-532.

[2] Peggy, O.O., 2025. Advanced machine learning-driven business analytics for real-time health risk stratification and cost prediction models. *World Journal of Advanced Research and Reviews*, *26*(2), pp.150-167.

[3] Akter, M.S., Sultana, N., Khan, M.A.R. and Mohiuddin, M., 2023. Business Intelligence-Driven Healthcare: Integrating Big Data and Machine Learning For Strategic Cost Reduction And Quality Care Delivery. *American Journal of Interdisciplinary Studies*, *4*(02), pp.01-28.

[4] Ok, E., 2024. How Predictive Analytics is Revolutionizing Financial Risk Management in Healthcare.

[5] Mehedy, M.T.J., Jalil, M.S., Saeed, M., Snigdha, E.Z., Khan, N. and Hasan, M.M., 2025. Big Data and Machine Learning in Healthcare: A Business Intelligence Approach for Cost Optimization and Service Improvement. *The American Journal of Medical Sciences and Pharmaceutical Research*, *7*(03), pp.115-135.

[6] Mabel, E. and Eniola, J., 2024. Data-Driven Financial Risk Management: The Role of Predictive Analytics in Healthcare.

[7] Markose, G.C., 2024. Predictive analytics for identifying high-risk Medicare patients: Enhancing preventive care. *Elevance Health Inc.[unofficial/preprint]*.

[8] Baiyewu, A.S., 2023. Overview of the role of data analytics in advancing health

service. *Open Access Library Journal*, *10*(6), pp.1-19.

[9] Mazumder, M.S.A., 2024. The transformative impact of big data in healthcare: Improving outcomes, safety, and efficiencies. *Global Mainstream Journal of Business, Economics, Development & Project Management*, *3*(03), pp.01-12.

[10] Dulam, A., 2025. Leveraging Data Analytics for Effective Risk Adjustment in the Affordable Care Act: Implications for Health Plan Management. *Journal of Computer Science and Technology Studies*, *7*(5), pp.290-297.

[11] Agarwal, D.P., Kushwaha, D.V., Singh, D.V.K., Azmi, D.T., Shukla, D.V., Khan, D.N.F. and Shoraisham, D.B., 2023. Revolutionizing Healthcare Through Advanced Analytics: Big Data.(2023). *Int J Pharm Sci*, *14*(4), pp.p62-74.

[12] Zainab, O.A. and Mgbole, T.J., 2024. Utilization of big data analytics to identify population health trends and optimize healthcare delivery system efficiency. *World Journal of Advanced Research and Reviews*.

[13] Badawy, M., Ramadan, N. and Hefny, H.A., 2023. Healthcare predictive analytics using machine learning and deep learning techniques: a survey. *Journal of Electrical Systems and Information Technology*, *10*(1), p.40.

[14] market.us, 2025, *Big Data in Healthcare Market*, Accessed On: 10th August, 2025, From: https://market.us/report/big-data-in-healthcare-market/

[15] Dash, S., Shakyawar, S.K., Sharma, M. and Kaushik, S., 2019. Big data in healthcare: management, analysis and future prospects. *Journal of big data*, *6*(1), pp.1-25.