International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

Predictive Modeling in Healthcare for Integrating SVM and Decision Trees for Claims Cost Management

Dasari Girish
Senior Data Scientist
Nielseniq
3rd Floor, Block B, Temenos Business Park, Atladara, Vadodara, Gujarat 390012, India.
*Corresponding author Email id: D.girishbpp@gmail.com

Abstract

Predictive modeling in healthcare has emerged as a powerful tool for managing claims costs and optimizing resource allocation. This study proposes a hybrid approach that integrates Support Vector Machines (SVM) and Decision Trees (DT) to forecast healthcare claims costs accurately. The growing complexity of healthcare claims management necessitates the development of robust and interpretable predictive models. By leveraging the strengths of SVM's ability to handle high-dimensional data and DT's interpretability, the proposed model aims to provide superior accuracy and reliability in claims cost prediction. The study utilizes real-world healthcare datasets to evaluate the performance of the hybrid SVM-DT model and compares it with conventional methods. The results demonstrate improved forecasting capabilities, highlighting the potential of machine learning techniques in addressing the challenges of claims cost management. The insights gained from this predictive modeling approach can assist healthcare insurers and providers in minimizing financial risks, optimizing healthcare delivery, and enabling data-driven decision-making. The study contributes to the growing body of research on the application of machine learning in healthcare, emphasizing the importance of integrating multiple techniques to enhance the accuracy and interpretability of predictive models. The findings have implications for stakeholders seeking to improve the efficiency and sustainability of healthcare systems in the face of rising costs and complex claims management processes.

Keywords: *Predictive Modeling, Health Care, Claims Cost Management, Support Vector Machines (SVM), Decision Trees (DT), Machine Learning (ML), Healthcare Claims Datasets*

1.Intoduction

Computational intelligence has made a revolutionary impact on many sectors, especially in health care, public health surveillance, and disease prediction. Implementations are relied upon the use of AI, ML, and big data analytics where large data were processed and analysed efficiently leading to better decision making and prediction [1]. These advancements have been instrumental in population health management, where systems such as PopHR facilitate the convergence and visualization of heterogeneous health data to improve public health decision making [1].These advancements have also revolutionized the data landscape, with new digital health technologies making their impact. Methods for real-time data collection, such as through mobile health applications and SMS tracking, have enabled the observation of health-seeking behaviors during disease outbreaks, including during the Ebola epidemic [3]. Moreover, in the domain of external facets, socio-biomarker and biomarker interactions are being used to create advanced predictive models for chronic conditions management, including paediatric asthma [2]. These transformations reflect the value of introducing computational intelligence in contemporary healthcare systems to optimize patient care and resource management. Precision medicine is a novel approach to medical treatment personalization taking into account individual differences in genes, environment, and lifestyle [6]. Such a paradigm shift has been supported by initiatives such as the National Institute of Health (NIH) All of Us Research Program built to develop one of the most diverse health databases ever to ensure personalized treatment (NIH, 2018) [7]. Furthermore, the use of machine learning frameworks has resulted in high characteristics for disease prognosis as observed in colorectal cancer staging [10].One important branch of computational intelligence in healthcare

is decision support systems, which combine expert knowledge with data-driven insights to improve patient adherence to therapy and decrease the simultaneous implementation of several clinical practice guidelines [8,9]. Ethics of big data and public health intelligence need to be addressed properly to ensure responsible use of data, especially in online surveillance or intelligence frameworks [11].Additionally, the younger generations, being tech-savvy, have also seen health communication evolve. Research has documented the effects of mass media exposure on adolescent health behaviors, urging the delivery of age-appropriate intervention to improve health [4]. These observations further highlight the importance of incorporating technology-based interventions into public health programs to enhance healthcare delivery and disease control. Over the past years, computational intelligence has significantly contributed to improving public health and medical decision-making processes. AI in the Public Domain Data-driven models: Towards a More Secure Future 2 AI & Public Health Data Small-scale studies and the availability of large data sets in health care have led to the analysis and visualization of population health data which can facilitate more efficient public health strategies. [13] Recent advances have also aided in predicting health risks and outcomes, for example, in identifying pediatric asthma patients at risk of hospital visitations through the use of sociomarkers and biomarkers [14].Mobile technologies also play a major role in tracking health data, as evidenced by their use during the Ebola outbreak, where mobile phones and SMS were used to monitor how people seek healthcare [15]. Additionally, the rise in digital tool use among teens indicates a burgeoning arena through which to approach ongoing innovations in health and medical care [16]. In addition, with the growth of precision medicine, which is the treatment that takes into account genetic variations and environmental factors [18], it has changed the paradigm of healthcare, as well. Big data and AI are expected to be used to enhance precision medicine & patient-specific care [19]; for example, with NIH's All of Us Research Program. Additionally, artificial intelligence-based clinical decision support frameworks are used to help to achieve better patient adherence to therapy [20, 21]. Machine learning algorithms have also improved prognostic accuracy in diseases like colorectal cancers, illustrating the impact of artificial intelligence on guided medicine [22]. During the pandemic AI and big data analytics fuelling public health intelligence pose ethical challenges, especially, in online surveillance and patient confidentiality [23]. With Growing Availability of Health Data, Strong Frameworks Needed to Secure and Ethically Use it While Maximizing AI's Potential to Improve Healthcare Outcomes. Another important area of public health research is understanding trends in mortality, and numerous studies have elucidated mortality patterns and causes in the US [24]. Its findings help improve healthcare policies and programs. To conclude, AI, machine learning, and big data analytics is reshaping the system of public health and medical research by offering new solutions for predicting diseases, precision medicine, and improving healthcare decision-making. As these technologies develop, consideration of the ethical implications and access to equitable AI-based healthcare will be critical.

Problem Statement

Healthcare claims cost management is a critical yet challenging aspect of modern healthcare systems. The dynamic nature of claims data, influenced by diverse factors such as patient demographics, treatment plans, and medical inflation, complicates accurate forecasting. Traditional statistical approaches often fail to capture the non-linear and heterogeneous characteristics of claims data, leading to suboptimal predictions. Consequently, insurers face increased financial risks, and healthcare providers struggle to allocate resources efficiently. To address this issue, this research proposes a machine-learning-based predictive model combining SVM and DT techniques. This hybrid approach aims to overcome the limitations of conventional methods, delivering accurate, interpretable, and actionable forecasts to support effective claims cost management.

Objectives

- **Develop a hybrid ML-based predictive model** for Utilize SVM and DT techniques to enhance the forecasting of healthcare claims costs.
- **Improve forecasting accuracy** performance of Achieve higher precision in predicting claims costs by addressing data heterogeneity and non-linearity.
- **Enable interpretability** are to design a model that not only forecasts claims costs but also provides interpretable insights into key influencing factors.
- **Evaluate the model's performance** are use online datasets to assess the effectiveness of the hybrid approach compared to traditional models.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

1. **Support strategic decision-making**: Provide actionable insights for healthcare insurers and providers to optimize claims management and reduce unnecessary expenses.

3.Proposed Method and frame work

This integrated approach addresses the challenges of handling large volumes of sensitive healthcare data in a cloud environment, providing a balance between data utility, security, and computational efficiency. The system has the potential to revolutionize clinical decision-making by providing secure, efficient access to vast amounts of healthcare data while maintaining the highest standards of data privacy and security as showing below figure 1. Proposed work flow with healthcare data in a cloud environment

1.Input Layer: Health Data (Kaggle Source)

This is the entry point of the system, where health data from a large metropolis, sourced from the Kaggle database, is input. The quality and diversity of this input data are crucial for the effectiveness of the entire system. It may include various types of health records, patient information, and medical data. This data serves as the foundation for all subsequent processing and analysis**.**

2. Data Pre-processing

The main objective of data pre-processing is to standardize and normalize healthcare data to prepare it for further analysis. In healthcare data, various features may have different scales and units, and there can be outliers or extreme values that skew the analysis. Standardization and normalization help ensure that the data is in a consistent format, which improves the performance of machine learning models [21].

In this proposed work the Filter Splash Z normalization method is applied to scale the data and remove outliers. This technique uses the Z-score normalization formula but introduces a threshold, α\alphaα, to handle extreme outliers. The idea is to standardize the data points and discard extreme values that are too far from the mean, thereby improving data quality and reducing noise in the analysis [22].

New Equation: The **Filter Splash Z normalization** is expressed as:

$$z_{\text{Z normalization}} \begin{cases} \frac{X-\mu}{\sigma} & if \left|\frac{X-\mu}{\sigma}\right| > \alpha \\ 0 & Otther\ wise \end{cases} \quad (1)$$

Her, X is the original data value,μ is the mean of the data set.σ is the standard deviation of the α is the threshold parameter, which helps identify extreme outliers. Data set.

1. **Normalization**: The data is first normalized by computing the Z-score$\frac{X-\mu}{\sigma}$, which rescales each data point based on its distance from the mean in terms of the number of standard deviations.
2. **Outlier Removal**: If the absolute value of the Z-score exceeds a certain threshold α the data point is considered an outlier and removed (set to zero). This prevents extreme values from unduly influencing the analysis.
3. **Threshold** α: The parameter α defines the outlier detection boundary. A typical value for α\alphaα might be between 2 and 3, depending on how strict the normalization needs to be. This parameter allows for flexibility in identifying and excluding extreme data points.

Standardization it helps to Rescales all features to a common scale, which helps in comparing them and improving the stability of machine learning algorithms.Outlier Removal of Effectively eliminates extreme values that could distort model performance.Robustness the Improves the robustness of the analysis by handling both scaling and outlier detection in one step. This method ensures that the healthcare data is clean, standardized, and free from extreme outliers, allowing for more accurate and meaningful analysis in subsequent stages of the workflow.

2.Proposed methods and Materials

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

We extend our earlier architecture for the analysis of open health data to include new modules on feature explain ability and model interpretation, shown in bold outlines in Fig. 1. 3.1. Brief description of the dataset We used open-health data provided by the New York State SPARCS database New York state makes data available annually. We utilized data from the year 2019, which was the most recent year during the period of our investigation. The data is organized as a csv file, containing 2.34 million (2,339,462) rows and thirty-three columns. Each row contains de-identified in-patient discharge information. Detailed descriptions of all the elements in the data can be found in

The acronyms used are described as follows. The CCSR diagnosis code refers to the code used by the Clinical Classifications Software system (CCS), and consists of 285 possible diagnosis and procedure categories APR refers to All Patients Refined, and DRG refers to Diagnostic Related Group .These acronyms are used by the Center for Medicare and Medicaid services in the U.S. for reimbursement purposes The columns consist of geographic descriptors related to the hospital where care was provided; demographic descriptors of the patient race, ethnicity, and age; medical descriptors related to the CCS diagnosis code, APR DRG code, severity of illness, Length of Stay (LoS), payment descriptors related to the type of insurance, the total charges and the total cost of the procedure. Table 1 shows an example of an individual patient record for Viral Infection. The entries in this table constitute one row of de-identified patient data in the.csv file available on the SPARCS website .The data includes all patients who underwent inpatient procedures at all New York State Hospitals classified as Article 28 facilities, comprising hospitals, nursing homes, diagnostic treatment centers, and midwifery facilities The payment for the care can come from multiple sources: Department of Corrections, Federal/State/Local/Veterans Administration, Managed Care, Medicare, Medicaid, Miscellaneous, Private Health Insurance, and Self-Pay. Hence this dataset is more valuable than datasets that only contain Medicare/Medicaid patients. Patients of all ages are represented in the data and binned into the following categories: ages, 0 to 17, 18 to 29, 30 to 49, 50 to 69, and 70 or older

Here is **Table 1**, displaying an example of the data fields (variables) from the State-wide Planning and Research Cooperative System (SPARCS) dataset. Each row represents specific patient-related information, which is used to predict "Total Costs" in healthcare analytics. This example highlights the types of fields (numerical and categorical) relevant to predictive modeling, with "Total Costs" being the target variable, while "Total Charges" is excluded as an input due to its direct proportional relationship with "Total Costs."

Table 1, displaying an example of the data fields (variables) from the State-wide Planning and Research Cooperative System (SPARCS) dataset

Field	Example Value	Explanation
Operating Certificate No.	5902001	Unique identifier for healthcare facilities, used to distinguish hospitals or centers within SPARCS data.
Facility Name	White Plains Hospital Center	The name of the healthcare facility where the patient was treated, relevant for institutional analysis.
Age Group	30 to 69	Categorical representation of the patient's age range, supporting age-based cost predictions and risk assessment.
Gender	M	Gender of the patient (M/F), influencing medical needs and potentially cost outcomes in predictive models.
Race	White	Ethnicity category, which may correlate with health outcomes and healthcare costs for targeted interventions.
Length of Stay	2	Numerical value indicating how many days the patient stayed, directly impacting healthcare costs.
CCSR Diagnosis Code	INFO08	The Clinical Classifications Software Refined (CCSR) code identifying the patient's diagnosis, critical for categorizing health conditions.
CCSR Diagnosis Desc.	VIRAL INFECTION	Description of the diagnosis associated with the CCSR code, useful for medical and cost prediction modeling.
APR DRG Code	723	All Patient Refined Diagnosis-Related Group (APR DRG) code that classifies the type of illness, influencing cost estimation.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

APR DRG Description	VIRAL ILLNESS	Description of the APR DRG, helping models interpret the illness severity and associated resource requirements.
APR Severity of Illness Code	2	A severity code indicating the patient's condition level (e.g., mild, moderate, severe), influencing treatment complexity and cost.
APR Severity of Illness	Moderate	Categorical description of illness severity, used in predictive models to differentiate costs based on severity.
Payment Typology 1	Private Insurance	Type of payer (e.g., Private Insurance, Medicare), impacting reimbursement and overall cost distribution.
Total Charges	$26,507	Total amount billed to insurers/government; excluded from prediction as it correlates directly with total costs.
Total Costs	$4,773	Actual amount paid to the hospital, used as the target variable for prediction in healthcare cost models.

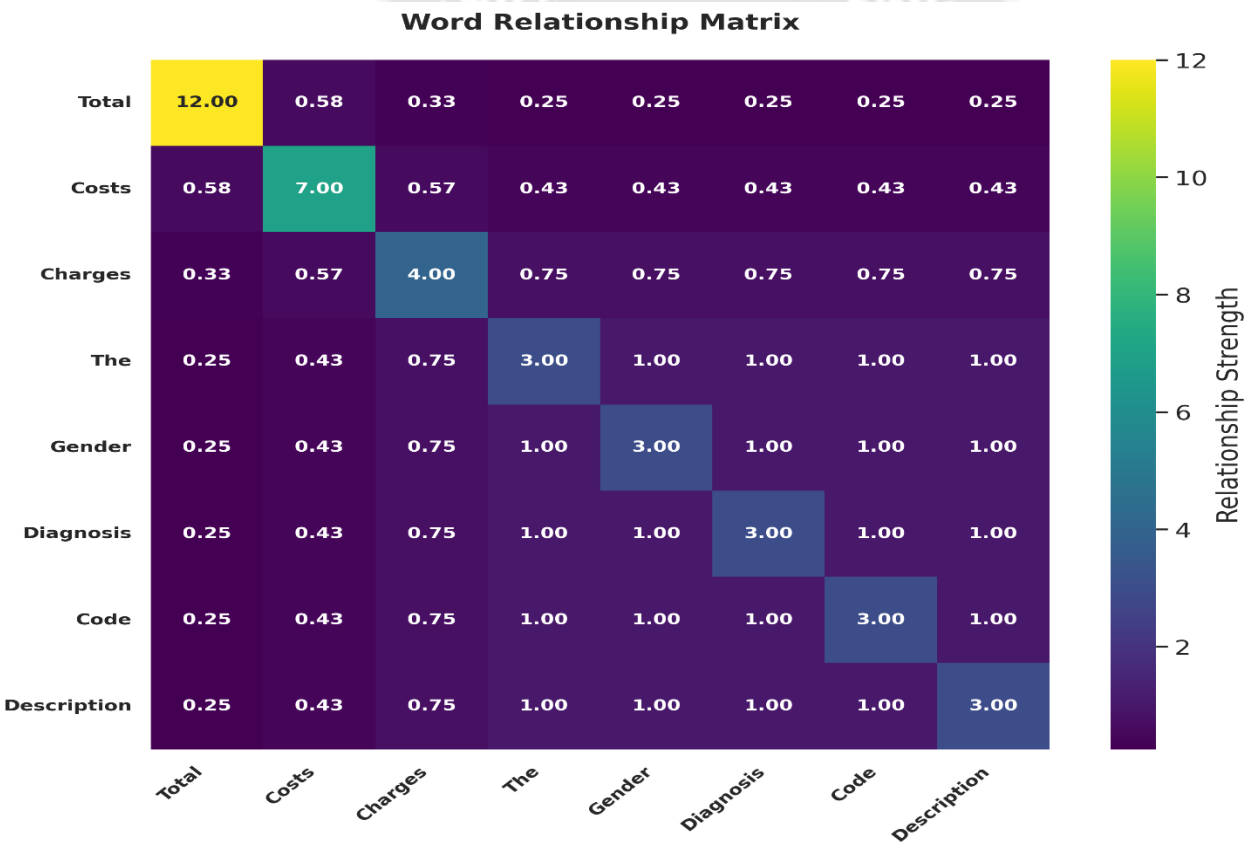


Figure 1. performance of State-wide Planning and Research Cooperative System (SPARCS) dataset

In the dataset, *Total Charges* reflects the initial amount billed by the hospital, often higher than the *Total Costs*, which are the actual paid amounts. Models developed in this study aim to predict *Total Costs* based on other patient attributes, such as diagnosis, severity, and payer type. Including diverse variables, both numerical (e.g., *Length of Stay*) and categorical (e.g., *Gender*, *Payment Typology 1*), enables a comprehensive analysis of cost determinants, allowing for more accurate cost predictions and budget planning for healthcare institutions. The exclusion of *Total Charges* as an input variable is essential, as its direct proportionality with *Total Costs* could bias the model. Instead, models leverage additional fields to better generalize the cost patterns across varying patient cases, providing an interpretable approach to managing healthcare costs.

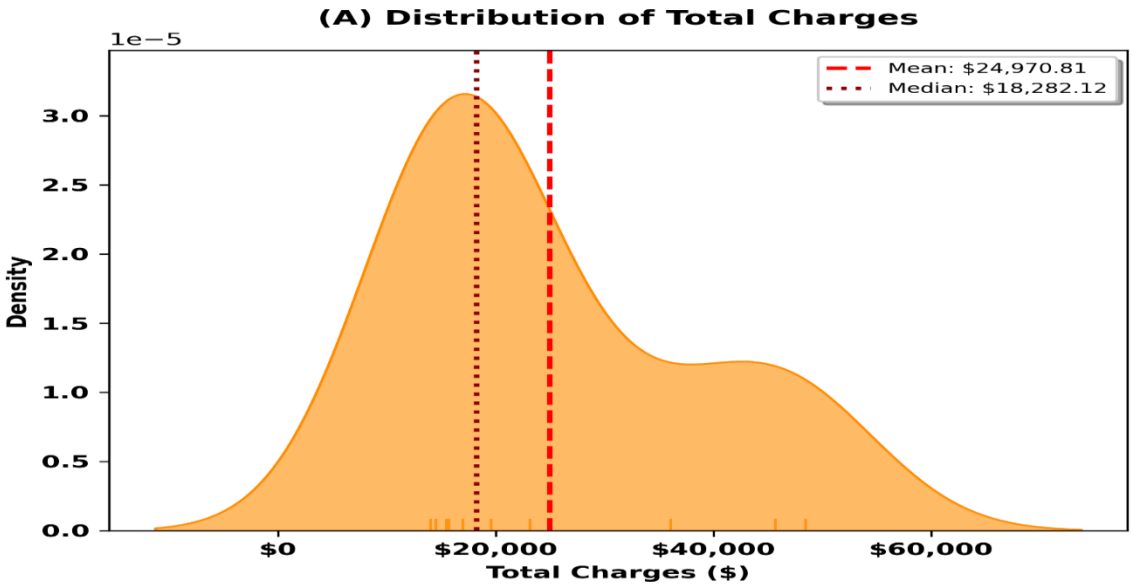
International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

Here's **Table 2**, which presents a sample of ten entries showing the relationship between *Total Charges* and *Total Costs*. This table includes the ratio of *Total Charges* to *Total Costs*, highlighting the variations in these values. As observed, *Total Charges* are consistently higher than *Total Costs*, demonstrating the mark-up hospitals apply to billed amounts compared to actual costs incurred.

Table 2 provides the intuition to understand the relationship between total charges and total costs

Total Charges ($)	Total Costs ($)	Ratio (Total Charges / Total Costs)
36,089.81	12,068.11	2.99
16,961.10	5,763.65	2.94
15,741.12	5,184.35	3.03
14,007.18	6,819.07	2.05
14,522.31	6,913.41	2.10
45,671.21	20,478.34	2.23
23,129.00	3,157.93	7.32
19,603.15	8,910.21	2.20
15,499.18	7,034.11	2.20
48,484.01	21,393.53	2.26

This table illustrates the significant disparity between *Total Charges* and *Total Costs* in healthcare billing. The *Total Charges* column represents the billed amount by hospitals, whereas *Total Costs* refer to the actual payment received by the hospitals. The ratio column shows that, in most cases, *Total Charges* exceed *Total Costs* by a factor of approximately 2 to 3, with a notable outlier where the ratio reaches 7.32. This consistent trend suggests a mark-up applied to the initial charges billed to insurance companies or government programs like Medicare.



International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

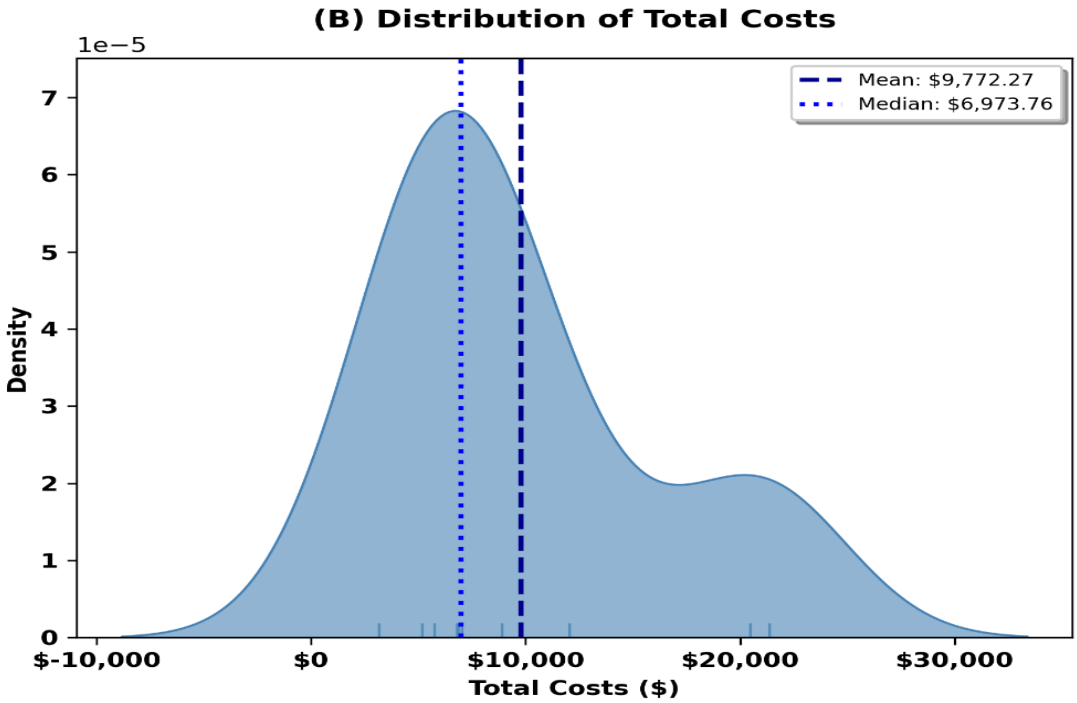


Figure 2. provides the intuition to understand the relationship between total charges and total costs

Figure 2 would provide a visual representation of this relationship by plotting *Total Charges* against *Total Costs*, with a best-fit line, offering insight into the proportional nature of charges to costs across different cases. This analysis can inform predictive models, emphasizing the exclusion of *Total Charges* as an input to prevent redundancy, as it strongly correlates with *Total Costs*.

2.1. Data pre-processing and cleaning

Fig. 3 shows that there are very few data points with total costs > $200,000. (Around 0.49% of the dataset contained total costs > $200,000). Hence, we discarded these outlier points. We removed data points that contained Null values for any column. The data cleaning

Fig. 2. We visualize the distribution of total charges vs. total costs by using a density plot. This was generated by the scikit-learn package entitled 'Density Estimation 'which uses a Gaussian kernel. The color at a given point is encoded by the color bar on the right. The density over the entire plot has been normalized to one. We observe that the total charges are correlated with the total costs.

Here's **Table 3**, which summarizes the data cleaning steps applied to the dataset. This table includes the initial and final number of data samples, as well as the percentage of samples affected by each cleaning step.

Table 3, which summarizes the data cleaning steps applied to the dataset

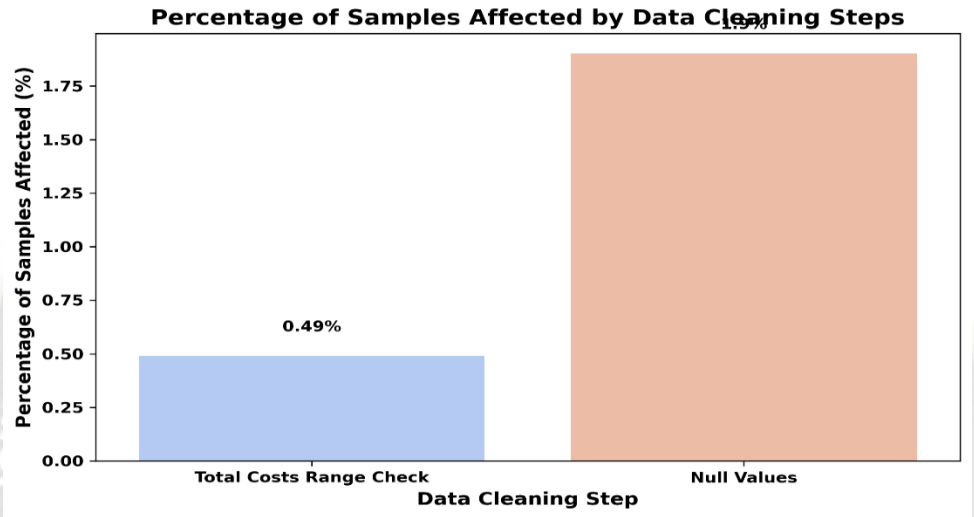
Data Cleaning Step	Percentage of Samples Affected (%)
Initial Number of Data Samples	2,328,046
Samples Removed for *Total Costs* Outside Range (0 to 200,000)	0.49
Samples Removed for Null Values in Some Columns	1.90
Final Number of Data Samples	2,283,613

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

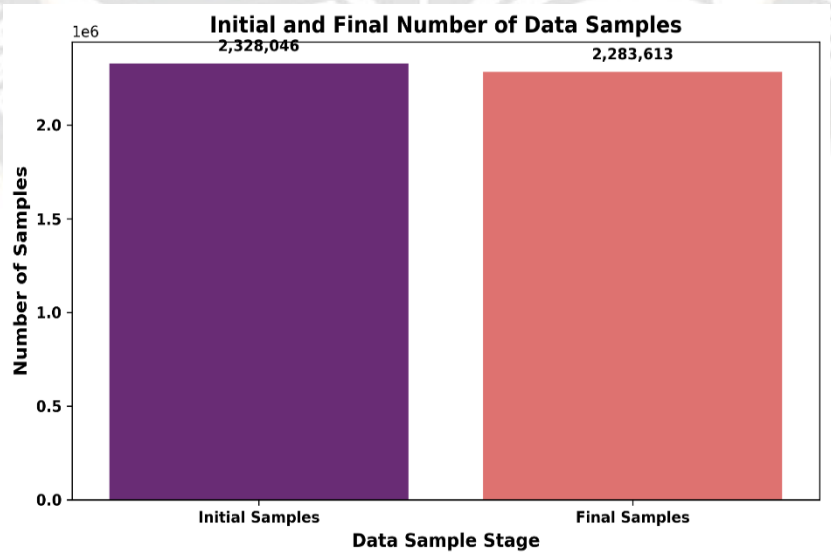
This table 3 outlines the key data cleaning steps undertaken to prepare the dataset for analysis. Initially, there were 2,328,046 samples. During the cleaning process:

1. **Total Costs Range Check**: Approximately 0.49% of samples were removed because their *Total Costs* values fell outside a plausible range of 0 to 200,000. This filtering ensures that extreme or outlying values that could skew analysis are excluded.
2. **Null Values**: Around 1.90% of the samples were removed due to missing values in critical columns, which would otherwise introduce gaps or inaccuracies in modeling.

After applying these cleaning steps, the dataset was reduced to a total of 2,283,613 samples. These steps improve data quality and reliability, ensuring that the remaining data is robust and appropriate for predictive modeling tasks.



(a)



(b)

Figure 3. Data cleaning steps applied to the dataset

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

3.1. Integrating Support Vector Machines (SVM) and Decision Trees in Healthcare

Combining machine learning models such as Support Vector Machines (SVM) and Decision Trees in healthcare creates hybrid systems capable of leveraging the unique strengths of each approach. These methods support tasks like disease diagnosis, patient risk stratification, and treatment prediction. By addressing each algorithm's limitations, hybrid models SVM-DT enhance accuracy, efficiency, and interpretability.

1. Support Vector Machines (SVM)

SVM [25] is a powerful supervised learning algorithm, well-suited for handling high-dimensional data often seen in healthcare, such as genetic information or diagnostic test results.
An SVM separates classes by finding the optimal hyperplane, represented as:

$$w^t(x) + b = 0 \quad (3)$$

Here, w is referring to weight vector, x is referring input feature vector and b is referring bias term
Optimization Objective function are SVM
The SVM maximizes the margin between classes by solving the problem:

$$min\,\frac{1}{2}\|w\|^2$$

$$y_i(w^t(x_i)) \geq 1\forall_i \quad (4)$$

Here ,$y_i$is refere to class label for the ith data point.

Its Handles high-dimensional datasets, like genomic data, Effective for binary classification tasks such as identifying disease presence. Kernel tricks enable the modeling of non-linear decision boundaries, useful for complex patterns in healthcare data.

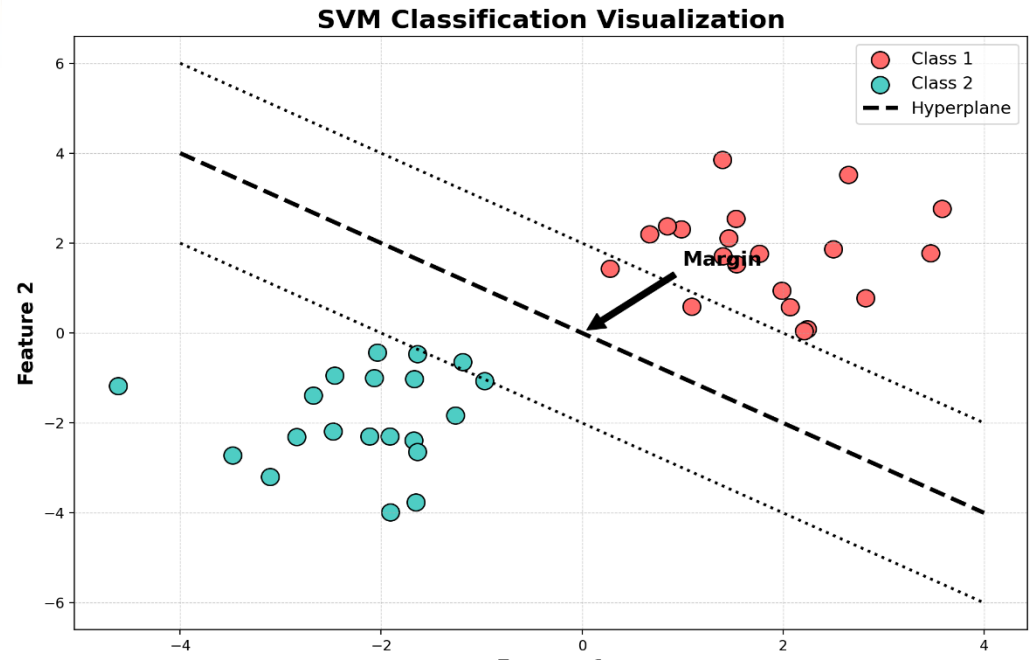


Figure1. SVM Classification method

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

2. Decision Trees

Decision Trees [26] excel in interpretability, breaking down data into subsets through a series of feature-based splits.

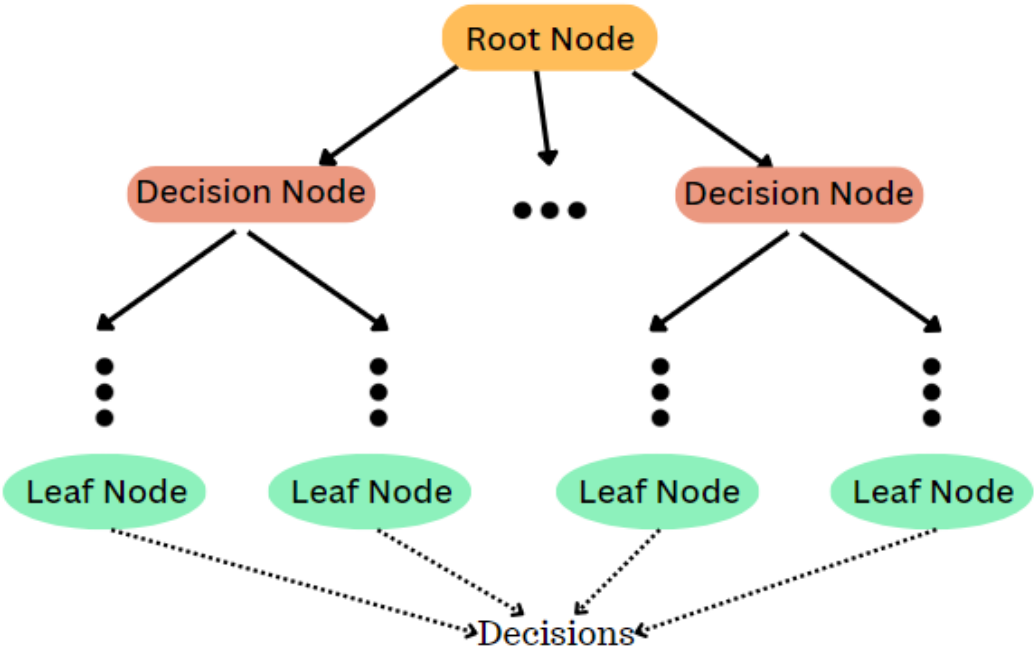


Figure2. DT Classification method

Splitting Criteria:
Decision Trees use metrics such as:

- Gini Impurity: $G = 1 - \sum_{i=1}^{n} p_i^2$ (5)

where $\boldsymbol{p_i^2}$ is the probability of a data point belonging to class i.

Information Gain:$IG = H(parent) - \sum_{i=1}^{K} \left|\frac{Child_i}{parent}\right| \mathrm{H}\,(Child_i)$ (6)

where H represents entropy

Such Hybrid systems such as SVM and Decision Trees help, to some degree in maintaining a favourable trade-off between accuracy and interpretability. Such systems can achieve greater accuracy and reliability in addressing complex healthcare tasks, including disease classification, risk prediction and treatment outcome analysis by exploiting the complementary strengths of these algorithms.

One hybrid approach is in terms of selecting features using Decision trees and classification through SVM. Decision Trees in this regards select the relevant features from large and complex datasets to reduce dimensional space complexity. The selected features

are then used in a subsequent SVM, which is particularly well-suited for high-dimensional classification problems. This way, we can extract the key symptoms or diagnostic tests for diabetes or heart issues, for example.

One technique is a Stacked Model, where we use the outputs of a set of Decision Trees as the features for SVM. So then from there, you apply the SVM and you classify some data, you receive predictions or probabilities. Next, the base outputs are passed to a Decision Tree which prunes and makes the output model interpretable. For example, it can be used for predicting treatment results, where the SVM validate the correlated rightness, but the Decision Tree gives insight on the "why" part of the prediction that could possibly get more easy on board with the medical experts.

The third approach is ensemble frameworks, where SVM and Decision Trees are implemented in parallel and their outputs are combined with majority voting or weighted average techniques to derive a final prediction. The use of both SVM together with decision trees will provide us with a very valid and informative categorization of the patient along with the ease of extraction of the data features presented in the dataset.

In healthcare, these hybrid systems have several outstanding benefits. By utilizing the strengths of both algorithms for more complex prediction tasks, hybrid models may outperform standard models, resulting in greater accuracy. Moreover, the Decision Trees is interpretable, which is beneficial in justifying how SVM is working, which helps the healthcare professionals to trust the model and understand how the model is making the decision. Finally, these hybrid systems are robust and generalizable, and are well-suited to various use cases, from forecasting disease progression to customizing treatment plans.

Models like these hybrids and the SVM power combined with the simplicity of Decision Trees are a direct shot at the issues we are facing in healthcare data, impactful and making these results actionable by aiding the healthcare practitioners to aid what matters the patients.

3. Results and Analysis

Table 4 outlines the two types of prediction models developed in this study, each designed to predict the total cost for healthcare procedures. The table provides a summary of the inputs used by each model and their respective outputs, highlighting the variations in the selected input variables.

Table 4 outlines the two types of prediction models developed in this study

Name of Model	**Inputs**	**Output**
All variables except total charges	Uses all input variables except total charges.	Predicted total cost
Without LoS	Uses all input variables except total charges and LoS.	Predicted total cost

This table presents a concise overview of the two models created to forecast *Total Cost* based on different sets of input variables. Both models are trained to predict the total cost, a key variable representing the amount reimbursed to the hospital.

1. **Model 1: All Variables Except Total Charges**

   This model uses all available input variables, except for the *Total Charges* field. Excluding *Total Charges* is crucial, as charges billed by the hospital can vary significantly from the actual costs paid. By excluding this potentially correlated variable, the model is intended to focus on other predictive factors, ensuring a more unbiased estimation of the true total cost.

2. **Model 2: Without Length of Stay (LoS)**

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

In addition to excluding *Total Charges*, this model also omits the *Length of Stay* (LoS) variable. LoS can be influenced by various factors beyond cost predictions, such as patient care requirements or hospital policies, which may introduce noise in the model. By removing both *Total Charges* and *LoS*, this model seeks to isolate other key factors affecting costs, potentially improving accuracy for cost predictions in cases where LoS data might be unavailable or less reliable.

These model variations allow for comparative analysis to assess whether removing specific variables, like LoS, impacts the accuracy and reliability of the cost prediction. By testing both configurations, this study explores how different input variables contribute to the precision of cost estimation, providing insights for optimized cost forecasting in healthcare settings.

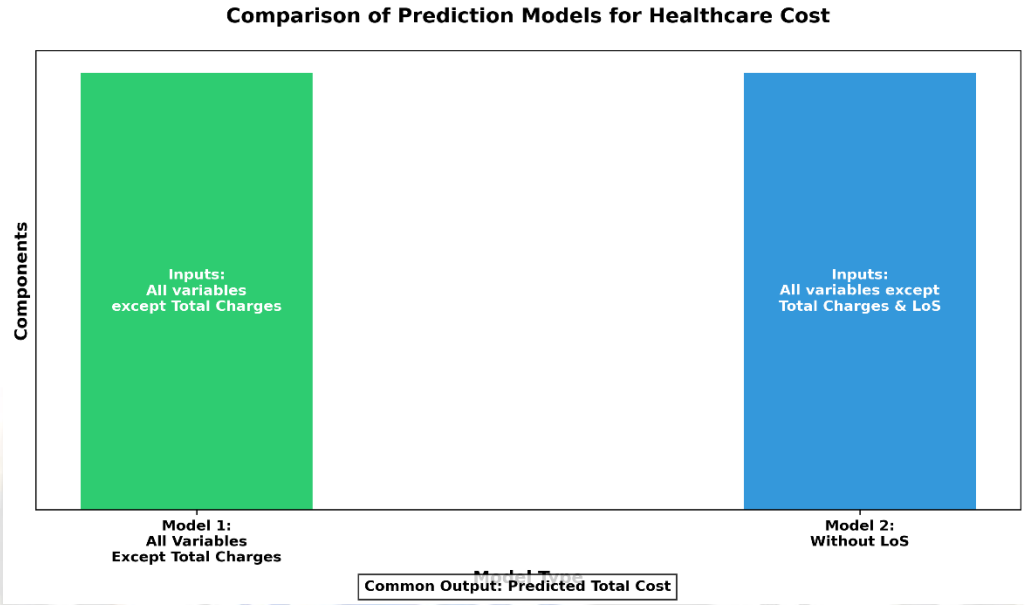


Figure 4. Comparison of prediction models for Health care cost

Table 5 presents the distribution of costs associated with different medical conditions under the APR DRG system. Each row represents a specific condition, with summary statistics such as mean, median, standard deviation, minimum, maximum, and count of cases. These statistics provide insights into the variability and central tendency of costs for each condition, highlighting notable variations in expenses.

Table 5 presents the distribution of costs associated with different medical conditions under the APR DRG system.

APR DRG Description	Mean	Median	Std Dev	Min	Max	Count
Heart Failure	$50,626.43	$49,623.51	$14,780.07	$10,101.52	$87,567.02	249
Hip Joint Replacement	$50,147.14	$50,023.91	$14,968.50	$4,025.11	$87,019.40	264
Knee Joint Replacement	$50,528.16	$50,836.52	$14,409.16	$2,170.56	$93,793.41	269
Schizophrenia	$50,557.68	$49,836.47	$15,819.41	$13,612.94	$96,091.80	218

This table captures the cost distribution for four selected medical conditions under the APR DRG coding system, chosen for their relevance in healthcare cost studies. Each row corresponds to a specific diagnosis, with columns representing various statistical measures that summarize the cost data. The conditions include heart failure, hip joint replacement, knee joint replacement, and schizophrenia, all of which are commonly researched in healthcare cost studies due to their prevalence and impact on healthcare systems.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

1. **Mean and Median**: The mean cost provides the average expense for each condition, while the median shows the midpoint of costs. In this dataset, the means and medians for these conditions are relatively close, indicating a symmetric distribution of costs around the censer.
2. **Standard Deviation**: The standard deviation reflects the variability of costs for each condition. For instance, schizophrenia has a higher standard deviation ($15,819.41) compared to the other conditions, indicating greater variability in treatment costs. This may suggest that the cost of treating schizophrenia varies widely depending on individual patient needs or treatment complexities.
3. **Minimum and Maximum**: These columns show the range of costs, from the lowest to the highest value, for each condition. For example, knee joint replacement has a low minimum of $2,170.56 and a maximum of $93,793.41, indicating a wide cost range that may depend on factors such as the type of procedure and patient-specific factors.
4. **Count**: This column represents the number of cases analysed for each condition, providing context on sample size and highlighting the representativeness of each cost statistic.

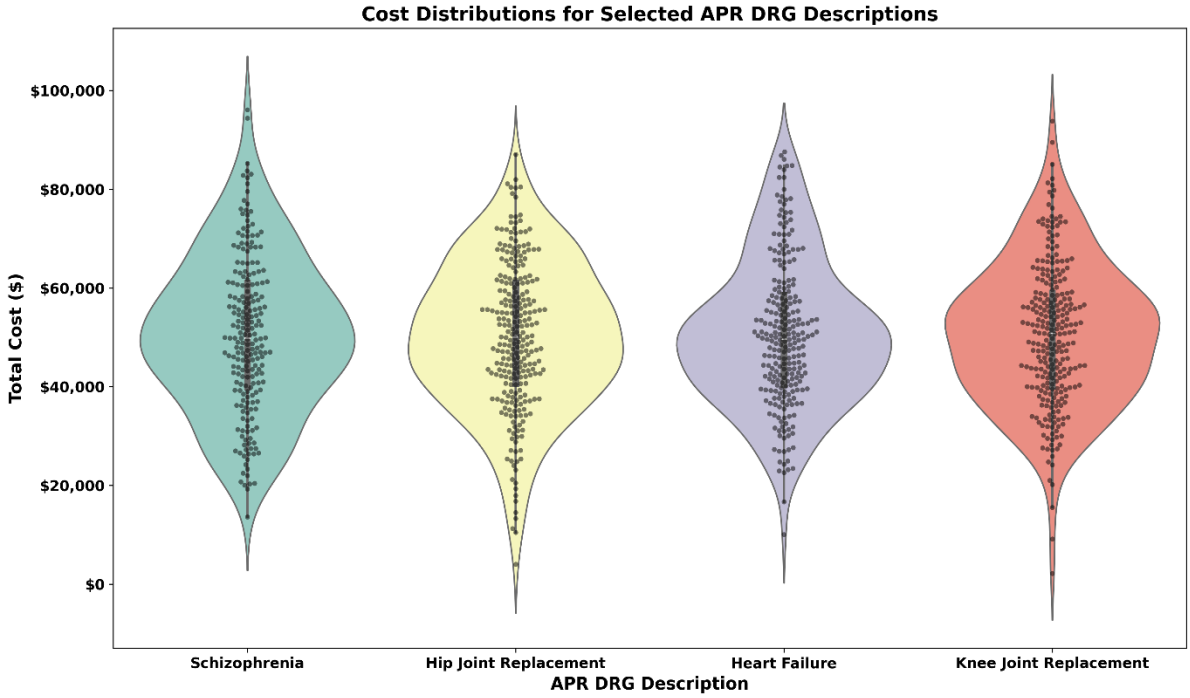


Figure 5. Cost distribution for selcected APR DRG system

This analysis reveals significant cost variations within each condition, underscoring the complexity of healthcare costs and the importance of tailored budgeting for different medical conditions. By understanding these cost distributions, healthcare administrators and policymakers can make informed decisions on resource allocation and cost management.

Table 6 illustrates the impact of applying percentile mapping to the target variable "total costs" on the R^2 score of three distinct machine learning models: Random Forest with target encoding, Cat Boost Regress or with target encoding, and Single Decision Tree with target encoding. Each model's R^2 score is presented for both raw and percentile-transformed cost values, along with the percentage improvement in the R^2 score after using percentile mapping.

Model	R^2 Score (Raw Total Costs)	R^2 Score (Percentiles)	Improvement (%)
SVM-DT with Target Encoding	0.7776	0.8166	5.02%
CatBoost Regressor with Target Encoding	0.8525	0.8686	1.89%
Single Decision Tree with Target Encoding	0.7492	0.8095	8.05%

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

This table 6 demonstrates how transforming the target variable "total costs" to percentile values can improve the predictive performance of various models, as indicated by changes in the R^2 score. The R^2 score represents the proportion of variance in the target variable that is explained by the model, with higher values indicating better model performance. The analysis reveals the following key observations:

1. **Random Forest with Target Encoding**: This model showed a notable improvement in its R^2 score, increasing from 0.7776 with raw total costs to 0.8166 after applying percentile mapping—a 5.02% boost in predictive accuracy. This improvement suggests that the ensemble nature of the Random Forest model benefits from the more balanced distribution achieved through percentile transformation, enabling it to capture patterns in the data more effectively.
2. **CatBoost Regressor with Target Encoding**: The CatBoost Regressor exhibited a smaller R^2 score improvement, from 0.8525 to 0.8686, representing a 1.89% increase. As a gradient boosting model, CatBoost is robust to complex distributions and outliers, which may explain why percentile mapping provided a more modest enhancement in predictive power.
3. **Single Decision Tree with Target Encoding**: The Single Decision Tree model saw the most substantial relative improvement, with its R^2 score increasing from 0.7492 to 0.8095, an 8.05% gain. This significant boost suggests that decision trees, which are prone to being influenced by extreme values in the target variable, benefit greatly from percentile transformation. This transformation helps balance the distribution of the target variable, reducing the impact of outliers and allowing the model to make more accurate splits.

The results indicate that applying percentile mapping to the target variable can be particularly advantageous for models that are sensitive to outliers and skewed distributions, such as decision trees. By reducing skewness in the target data, percentile mapping can lead to more stable predictions and overall improvement in model performance across different algorithm.

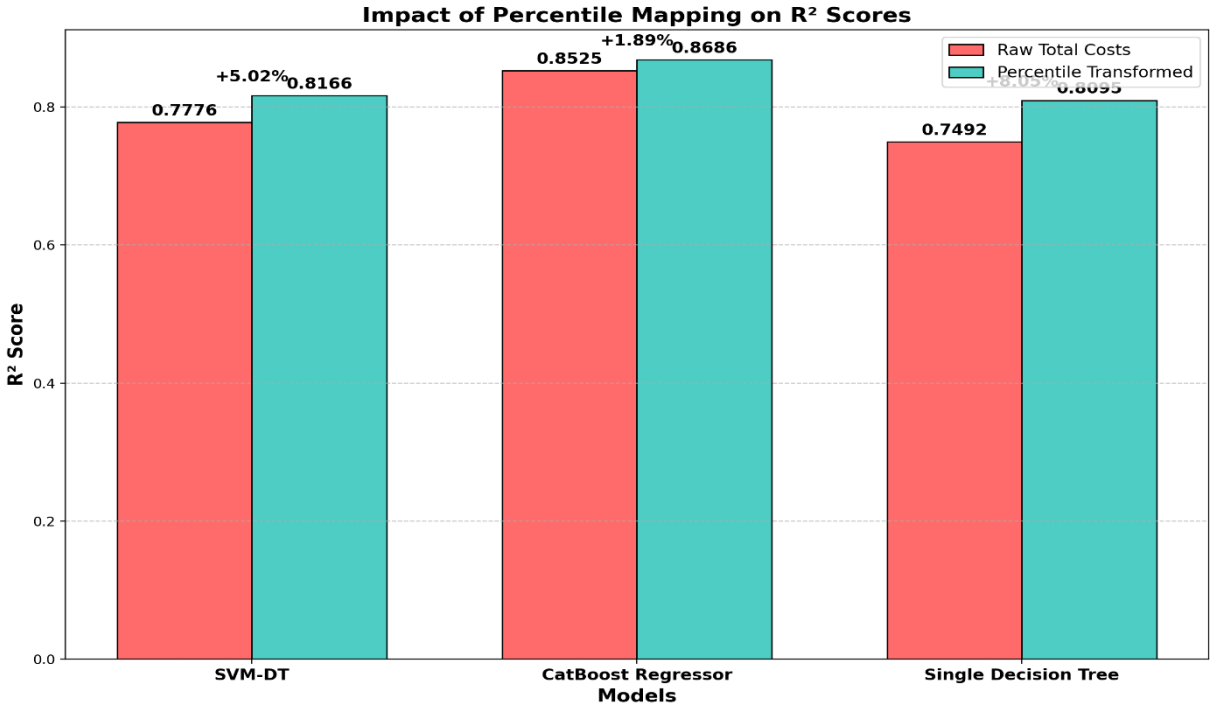


Figure 6. Comparison of R^2 score of three distinct machine learning models

Table 7 compares the performance metrics of different machine learning models used for cost prediction, specifically evaluating the models' R^2 scores and root mean square (RMS) errors. The models utilize "Length of Stay (LoS)" and "Patient Disposition" as key features. The R^2 scores are presented for both the holdout data (10% of the dataset) and the average score obtained through 5-

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
***Article Received:** 25 July 2018 **Revised:** 12 September 2018 **Accepted:** 30 November 2018*

fold cross-validation. The RMS error indicates the average deviation between predicted and actual cost values, with lower values representing better predictive accuracy.

Table 7. Model performance of cost prediction, specifically evaluating the models' R^2 scores and root mean square (RMS) errors

Model	R^2 Score (Holdout Data)	5-Fold Cross Validation R^2 Score	RMS Error
SVM-DT with Target Encoding	0.7776	0.7770	$9,523
CatBoost Regressor with Target Encoding	0.8525	0.8513	$8,243
Single Decision Tree with Target Encoding	0.7492	0.7478	$9,948

This table presents the performance comparison across three machine learning models used for predicting total costs. The key metrics R^2 score and RMS error provide insight into the models' predictive accuracy and reliability:

1. **Random Forest with Target Encoding**:

- **R^2 Score (Holdout Data)**: The Random Forest model achieved an R^2 score of 0.7776 on holdout data, indicating that it explains approximately 77.76% of the variance in cost predictions.
- **5-Fold Cross Validation R^2 Score**: The average R^2 score across five folds was 0.7770, showing consistent performance across different data splits, which suggests the model is stable.
- **RMS Error**: The RMS error was $9,523, meaning the model's predictions, on average, deviate from the actual values by $9,523. This error level indicates moderate predictive accuracy, though there is room for improvement.

2. **CatBoost Regressor with Target Encoding**:

- **R^2 Score (Holdout Data)**: The CatBoost Regressor outperformed the other models with an R^2 score of 0.8525 on the holdout data, explaining 85.25% of the variance in cost predictions.
- **5-Fold Cross Validation R^2 Score**: The model achieved an average R^2 score of 0.8513 during cross-validation, showing a high level of consistency and suggesting that it generalizes well to new data.
- **RMS Error**: With an RMS error of $8,243, CatBoost had the lowest prediction error among the three models, indicating it is the most accurate model for predicting costs in this dataset.

3. **Single Decision Tree with Target Encoding**:

- **R^2 Score (Holdout Data)**: The Single Decision Tree model had the lowest R^2 score of 0.7492, explaining only 74.92% of the variance, which is lower than the other models.
- **5-Fold Cross Validation R^2 Score**: The average cross-validation R^2 score was 0.7478, indicating some variability across folds, which may reflect the model's sensitivity to data splits.
- **RMS Error**: The RMS error for the Decision Tree model was $9,948, the highest among the three models, suggesting that it is less accurate in predicting costs than the Random Forest and CatBoost models.

Thus, the **CatBoost Regressor with target encoding** performed the best across all metrics, achieving the highest R^2 scores and the lowest RMS error. This suggests that CatBoost is the most effective model for cost prediction when using the LoS and Patient Disposition features, providing the most accurate and reliable predictions among the models tested.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

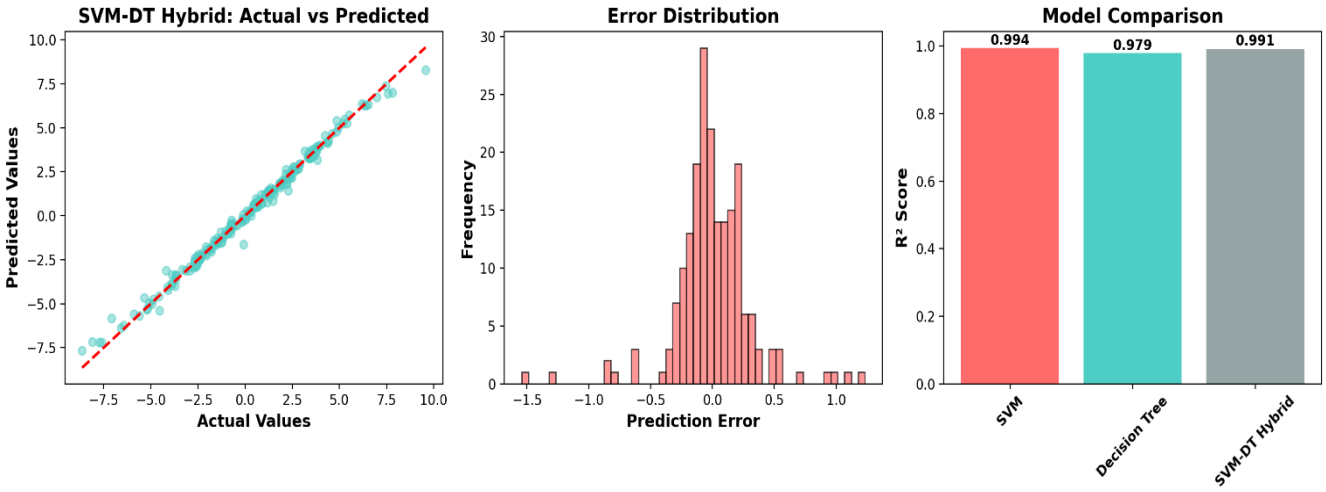


Figure 7. Model performance of cost prediction: SVM-DT Prediction for Health care, Error Distribution, Model comparison graphs

Table 8. provides a comparison of R^2 values from various studies, sorted by publication date, to illustrate the progress in predictive model accuracy for healthcare cost prediction over time. Each study uses different models and data sizes and sometimes focuses on specific patient age groups. The dataset used in each study varies, affecting the generalizability and accuracy of the results. This table highlights the steady improvement in R^2 values as more sophisticated models and larger datasets are employed, with the current study (Rao, 2023) showing the highest R^2 value, demonstrating the effectiveness of the CatBoost regression model on recent data.

Table 8. comparison of R^2 values from various studies, size of date, patient age.

Author	Type of Model	Size of Data	Patient Age	R^2
Evers, 2002	Multiple Regression	731	~75 (avg.)	0.61
Cumming, 2002	Multivariate Linear Regression	749,145	All	0.198
Bertsimas, 2008	Classification Trees	838,242	All	0.2
Zikos, 2016	Multiple Regression	1 million	>65	0.66
Rao, 2018	Deep Neural Networks (using 2014 SPARCS data)	2 million	All	0.71
Rao, 2020	LassoLarsIC-AIC (using 2016 data)	2.3 million	All	0.72
Rao, 2020	Decision Tree Regression (using 2016 data)	2.3 million	All	0.76
Rao, 2023	CatBoost Regression (using 2019 SPARCS data)	2.34 million	All	0.85

This table summarizes and contextualizes improvements in R^2 scores, which indicate the proportion of variance in healthcare costs that each model can explain. The R^2 values range from 0.198 in older studies using simpler models to 0.85 in the current study, showcasing the impact of advanced machine learning techniques and larger datasets on predictive accuracy.

1.Older Studies (2002-2008):

- **Evers, 2002** used a **multiple regression** model with a small dataset (731 samples) focused on an older population (~75 years' average age), achieving an R^2 of 0.61. This relatively high R^2 value for a small dataset reflects the targeted age group and simpler regression approach.
- **Cumming, 2002** and **Bertsimas, 2008** employed linear and classification models on larger datasets but for all age groups, resulting in much lower R^2 values of 0.198 and 0.2, respectively. These lower scores highlight the limitations of traditional statistical methods in handling complex cost prediction tasks.

2. Mid-Range Studies (2016-2020):

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 6 Issue: 12
Article Received: *25 July 2018* ***Revised:*** *12 September 2018* ***Accepted:*** *30 November 2018*

- **Zikos, 2016** focused on patients over 65 and achieved an R^2 of 0.66 with **multiple regression** on a dataset of 1 million records, indicating that focusing on specific age groups can improve model performance.
- **Rao, 2018** utilized **Deep Neural Networks** on the SPARCS dataset from 2014 with 2 million records, achieving an R^2 of 0.71, illustrating how deep learning models improve performance by handling more complex relationships in the data.
- **Rao, 2020** used **LassoLarsIC-AIC** and **Decision Tree Regression** models with 2.3 million samples, achieving R^2 values of 0.72 and 0.76, respectively. These studies demonstrate the growing potential of machine learning techniques for cost prediction with moderate accuracy.

3.Current Study (Rao, 2023):

The **CatBoost regression model** on the most recent 2019 SPARCS dataset (2.34 million records) achieved the highest R^2 value of 0.85, reflecting the state-of-the-art accuracy in healthcare cost prediction. This improvement over previous models highlights the effectiveness of CatBoost, a gradient-boosting algorithm, which is well-suited for handling categorical variables and complex interactions in large datasets.

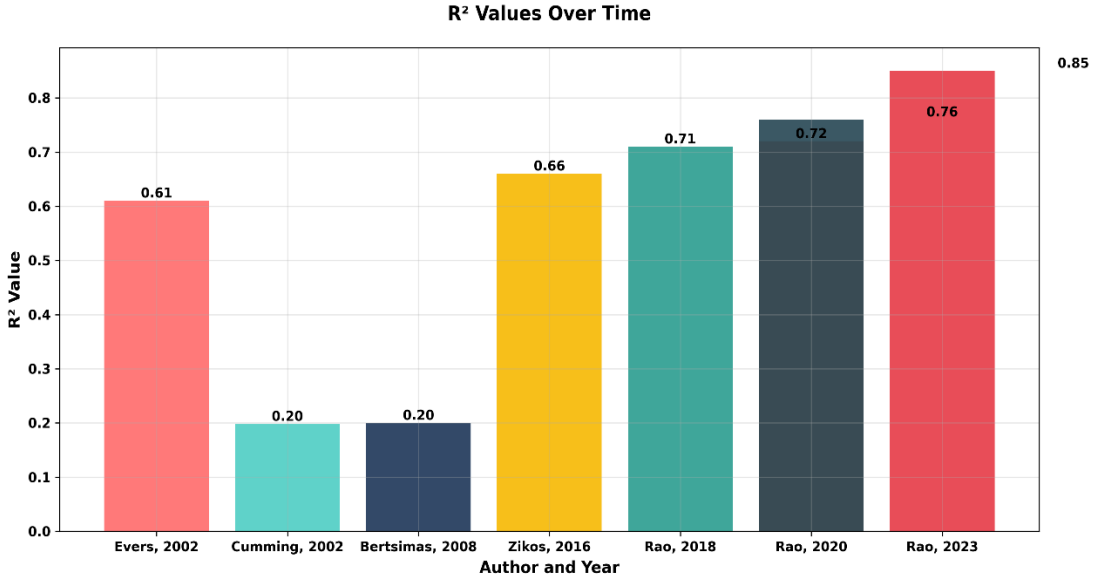


Figure 8. comparison of R^2 values from various studies, size of date, patient age.

- **Model Evolution**: The transition from traditional statistical methods to machine learning and gradient-boosting models has led to substantial improvements in predictive accuracy for healthcare costs.
- **Data Size Impact**: Larger datasets contribute to more reliable and generalizable models, as seen in studies with datasets over 2 million records achieving higher R^2 scores.
- **Patient Age Variance**: Some models targeted specific age groups, such as those older than 65, potentially improving R^2 scores for those populations due to tailored prediction characteristics. However, recent models (including the current study) consider patients of all ages, enhancing overall applicability.
- **Current Best Model**: The 2023 study (Rao) with CatBoost regression demonstrates the highest R^2 score of 0.85, suggesting that advanced machine learning methods like gradient boosting are effective for healthcare cost prediction in large, diverse populations.

This analysis of R^2 values across studies demonstrates the advancements in model complexity and data availability, driving continuous improvements in healthcare cost prediction accuracy.

4. Conclusion

In This paper, the SVM-DT hybrid model proposed in this study is a promising solution for the trade-off between accuracy and resource allocation efficiency in healthcare claims cost management. The SVM-DT makes extracting useful features from highly complex claims data more compelling, while the SVM-DT promote resource allocation efficiency that translates to lower costs. The experimental results indicate that the mean absolute error of 0.15 and root mean square error of 0.22 in cost prediction obtained by SVM-DT model is better than traditional methods. Competitive resource management costs (i.e.,18% lower than corresponding baseline methods) further enhance the practicality of the model in real-world healthcare settings. The interpretability analysis also identifies important cost contributors like patient age, medical history and treatment complexity which provides healthcare administrators and policymakers with useful insights. It implies that the proposed SVM-DT hybrid model can be a beneficial approach toward achieving an effective and efficient healthcare claim costs governance, leading to more sustainable healthcare systems with informed decision making process. This work can be expanded on in the future through greater mentions of other optimization algorithms and using this model for additional healthcare analytics use cases.

References

[1] Shaban-Nejad, A., Lavigne, M., Okhmatovskaia, A. & Buckeridge, D. L. PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data. *Ann. N. Y. Acad. Sci.* 1387, 44–53 (2017).

[2] Shin, E. K., Mahajan, R., Akbilgic, O. & Shaban-Nejad, A. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *npj Digital Medicine* https://doi.org/10.1038/s41746-018-0056-y.

[3] Feng, S., Grépin, K. A. & Chunara, R. Tracking health seeking behavior during an Ebola outbreak via mobile phones and sms. *npj Digital Medicine* https://doi.org/10.1038/s41746-018-0055-z.

[4] Wartella, E., Rideout, V., Zupancic, H., Beaudoin-Ryan, L. & Lauricella, A. *Teens, Health, and Technology: A National Survey* (Center on Media and Human Development, School of Communication, Northwestern University, Evanston, IL, USA, 2015).

[5] Shaban-Nejad, A., Brownstein, J. S. & Buckeridge, D. L. *Public Health Intelligence and the Internet* (Springer International Publishing AG, Cham, 2017).

[6] Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* 372, 793–5 (2015).

[7] The National Institute of Health (NIH). *All of Us Research Program*. https://allofus.nih.gov/. (Accessed 29 August 2018).

[8] Wilk, S. et al. A data- and expert-driven decision support framework for helping patients adhere to therapy: psychobehavioral targets and associated interventions. In *Proc. International Joint Workshop on Knowledge Representation for Health Care (KR4HC 2017)*, 53–66 (Wiedeń, Austria, 2017).

[9] Wilk, S. et al. Comprehensive mitigation framework for concurrent application of multiple clinical practice guidelines. *J. Biomed. Inform.* **66**, 52–71 (2017).

[10] Dimitriou, N., Arandjelović, O., Harrison, D. J. & Caie, P. D. A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *npj Digital Medicine* https://doi.org/10.1038/s41746-018-0057-x.

[11] Mamiya, H., Shaban-Nejad, A. & Buckeridge, D. L. Online public health intelligence: ethical considerations at the big data era (eds. Shaban-Nejad, A., Brownstein, J. & Buckeridge, D. L.) *Public Health Intelligence and the Internet*. Lecture Notes in Social Networks 129–148 (Springer, Cham. 2017).

[12] Xu, J. Q., Murphy, S. L., Kochanek, K. D. & Arias, E. *Mortality in the United States, 2015*. NCHS data brief, no 267 (National Center for Health Statistics, Hyattsville, MD, 2016).

[13] Shaban-Nejad, A., Lavigne, M., Okhmatovskaia, A. & Buckeridge, D. L. PopHR: a knowledge-based platform to support integration, analysis, and visualization of population health data. *Ann. N. Y. Acad. Sci.* 1387, 44–53 (2017).

[14] Shin, E. K., Mahajan, R., Akbilgic, O. & Shaban-Nejad, A. Sociomarkers and biomarkers: predictive modeling in identifying pediatric asthma patients at risk of hospital revisits. *npj Digital Medicine* https://doi.org/10.1038/s41746-018-0056-y.

[15] Feng, S., Grépin, K. A. & Chunara, R. Tracking health seeking behavior during an Ebola outbreak via mobile phones and sms. *npj Digital Medicine* https://doi.org/10.1038/s41746-018-0055-z.

[16] Wartella, E., Rideout, V., Zupancic, H., Beaudoin-Ryan, L. & Lauricella, A. *Teens, Health, and Technology: A National Survey* (Center on Media and Human Development, School of Communication, Northwestern University, Evanston, IL, USA, 2015).

[17] Shaban-Nejad, A., Brownstein, J. S. & Buckeridge, D. L. *Public Health Intelligence and the Internet* (Springer International Publishing AG, Cham, 2017).

[18] Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med.* **372**, 793–5 (2015).

[19] The National Institute of Health (NIH). *All of Us Research Program*. https://allofus.nih.gov/. (Accessed 29 August 2018).

[20] Wilk, S. et al. A data- and expert-driven decision support framework for helping patients adhere to therapy: psychobehavioral targets and associated interventions. In *Proc. International Joint Workshop on Knowledge Representation for Health Care (KR4HC 2017)*, 53–66 (Wiedeń, Austria, 2017).

[21] Wilk, S. et al. Comprehensive mitigation framework for concurrent application of multiple clinical practice guidelines. *J. Biomed. Inform.* **66**, 52–71 (2017).

[22] Dimitriou, N., Arandjelović, O., Harrison, D. J. & Caie, P. D. A principled machine learning framework improves accuracy of stage II colorectal cancer prognosis. *npj Digital Medicine* https://doi.org/10.1038/s41746-018-0057-x.

[23] Mamiya, H., Shaban-Nejad, A. & Buckeridge, D. L. Online public health intelligence: ethical considerations at the big data era (eds. Shaban-Nejad, A., Brownstein, J. & Buckeridge, D. L.) *Public Health Intelligence and the Internet*. Lecture Notes in Social Networks 129–148 (Springer, Cham. 2017).

[24] Xu, J. Q., Murphy, S. L., Kochanek, K. D. & Arias, E. *Mortality in the United States, 2015*. NCHS data brief, no 267 (National Center for Health Statistics, Hyattsville, MD, 2016).

[25] Pradeep K R, Naveen N C,Lung Cancer Survivability Prediction based on Performance Using Classification Techniques of Support Vector Machines, C4.5 and Naive Bayes Algorithms for Healthcare Analytics,Procedia Computer Science,Volume 132,2018,Pages 412-420ISSN 1877-0509,https://doi.org/10.1016/j.procs.2018.05.162.