10.48047/jocaaa.2023.31.04.26

AI-Driven Predictive Analytics in Healthcare: Evaluating Impact on Cost and Efficiency

Md Nazmul Alam Bhuiyan

MBA in Data Analytics

University of New Haven, CT, USA

mbhui2@unh.newhaven.edu

bhuiyan.unh@gmail.comS

0009-0003-2830-3396

Rabi Sankar Mondal

MSc. in Business Analytics

University of New Haven, CT, USA

rmond1@unh.newhaven.edu

rabi.s.mondal@gmail.com

0009-0006-0136-9354

Abstract

With the increased cost of healthcare and operational efficiency under the microscope, deploying intelligent decision-support systems has become a prerequisite. This study introduces and discusses a two-pronged AI framework that is conceptualised to address two significant challenges: Accurate costing of the population's healthcare and partitioning of the population based on their cost-effectiveness. Using a set of 1,338 healthcare insurance records, the study employs a regression and classification approach to comprehend healthcare spending and the ability to forecast resource allocation difficulties. Cost prediction is achieved using Linear Regression, Random Forest Regressor, and XGBoost Regressor, where Random Forest achieves the best performance ($R^2 = 0.87$; RMSE ≈ 4568). To categorise the patients, this research employs Logistic Regression, Random Forest Classifier, and XGBoost Classifier to divide them into the low, medium, and high-cost groups, and Random Forest outperforms with a high macro F1-score (0.8953) and accuracy (89.93 Apart from its favourable predictive results, the method helps to make budgeting decisions, sets the care hierarchy, and establishes policies. The union of transparent differential algorithms for ensembles with strategic healthcare goals can provide the correlation between financial projections and the optimisation of that clinical resource allocation. The framework is scalable and interpretable and can be easily applied in hospital settings, insurance systems, and public health information systems.

Keywords: Predictive Analytics**,** Machine Learning in Healthcare**,** Medical Cost Prediction**,** Patient Stratification**,** Random Forest and XGBoost, Logistic Regression

1 Introduction

Worldwide healthcare systems endure heavy pressures from rising costs, increasing complexity of patients, and insufficient distribution of resources. According to data presented by the World Health Organisation and several national bodies, soaring medical expenses are outstripping the GDP growth in many countries, threatening the viability of public and private facilities regarding health services [1]. Preventive admissions, misdirected patient flow, and difficulty in predicting resource demands

10.48047/jocaaa.2023.31.04.26

represent a significant proportion of the total costs. Moreover, problems like billing inefficiencies, manual triage, and reactive care models naturally burden finances. Such issues threaten healthcare providers, payers, and the overall standard of care patients receive. In such environments, data-driven approaches capable of anticipating financial stress and roadblocks to operations are becoming increasingly important. It is possible to forecast both patient and system-scaled dynamics using predictive modelling, allowing healthcare organisations to shift from reactive to proactive approaches towards cost and efficiency management.

The use of Artificial Intelligence (AI), including machine learning (ML) as a driver of change in healthcare analytics, is increasingly becoming clear. Through the use of structured and semi-structured data, artificial intelligence-driven predictive analytics uncovers insights, as well as trends that are often irrelevant using traditional statistical methods. In regards to financial and operational efficiency, machine learning (ML) models can predict individual patients' outcomes such as overall hospital cost, length of stay, and readmission probability, and classify patients on cost implications and care needs [2]. Such models as Random Forest and the XGBoost have shown pronounced capacities to learn complex, as opposed to linear, clinical and administrative data dependencies. Added diversity of metrics like age, BMI, region, businesses, and lifestyles enhances the cost forecasting and operational planning efficiency. With the ready digital health records and insurance claim data, the introduction of ML in the financial model of risk and operational forecasting is now imperative.

Despite the expanding role of AI in healthcare, most implementations focus on either regression-based financial forecasting or classification-based clinical outcomes, but rarely both. Various practices in current cost modelling favour linear regression, although this is insufficient to describe the complex, non-linear relations within actual healthcare datasets [3]. By contrast, the characteristics of classification models are that they are used as segmentation tools for patients in terms of clinical risk rather than for economic optimisation. The existing tools are limited because they cannot combine the evaluation of continuous regression to evaluate cost prediction with multiclass classification to segment operational risk. There is a demand for procedures that assist healthcare providers and insurers in predicting individual cost profiles and classifying patients into operational risk groups with low, medium, or high resource intensity. The lack of a dual-track AI model results in incomplete understanding, less effective planning and a lack of the opportunity to make strategic use of it.

This research attempts to fill these gaps by developing and evaluating a dual-track AI framework that should be utilised in cost and efficiency modelling. This study will use regression models such as linear, random forest, and XGBoost regression to forecast individual healthcare charges to attain the first objective. The second goal is to segment patients based on their costs: low, medium, or high, using classification techniques such as Logistic Regression, Random Forest Classifier and XGBoost Classifier. The third objective involves the evaluation of model performance on a set of metrics, including Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2 score for regression, and Accuracy, Precision, Recall, F1-score, Confusion Matrix, and ROC-AUC for classification. When the research accomplishes these objectives, it provides a reliable and easy method for healthcare administrators to forecast costs and increase efficiency.

2 Literature Review

2.1 Predictive Analytics in Healthcare: Key Trends and Past Studies

Predictive analytics has experienced significant development and popularity in healthcare organisations in the last two decades. Due to the growing use of electronic health records, insurance claims databases and patient registries, healthcare institutions have turned to more data-driven strategies for decision-making to reach better results and savings [4]. In healthcare, predictive modelling denotes the use of historical data and the application of statistical or machine learning methods to estimate such forthcoming events as readmissions, mortality rates, a period of hospitalisation, or financial costs. Many studies have revealed that predictive analytics are indeed helpful in making healthcare and office decisions. As [5] deep learning analysis of electronic health records resulted in a substantial increase in

10.48047/jocaaa.2023.31.04.26

the accuracy of predicting several clinical events compared to standard models. [6] also developed predictive models that strive to identify patients in the earlier stages of their care who may benefit most from palliative care, showing how AI can improve quality without the burden of unnecessary procedures. In the field of finance, there is a growing application of predictive systems for identifying top-cost contributors for prevention interventions in care. This change shows that more people now feel that predictive analytics can become helpful in healthcare, with clinical and non-clinical means, such as operations and finances.

2.2 Cost Modelling and Stratification: Statistical vs ML Approaches

Linear regression and actuarial methods have conventionally been used in healthcare cost modelling with assumptions of normality, independence, and linearity, confining it [3]. Conventional methods take patient demographics and medical histories to estimate future healthcare expenditure or service use. However, their predictive potential is limited because they cannot adequately model the complex interacting processes, non-linear relations, and non-Gaussian shocks frequently observed in health care spending. Due to improvements in machine learning, it is now possible to construct more flexible models that go beyond the limitations present in earlier approaches. Such models as Random Forest, Gradient Boosting Machines, and XGBoost are very good for capturing intricate interactions in big datasets, and they outperformed the traditional methods in cost estimations. Generalised linear models were typically outperformed by tree-based models in estimating hospital charges, according to [7]. Both [8, 9] have shown that the ensemble models led to higher R. ML-based approaches increase prediction capabilities and enable easy incorporation in contemporary digital systems in real time, thus fitting very well with agile healthcare and insurance operations. Although improvements exist, a unified system has not fully realised the synergistic use of cost–modelling and clinical classification methods.

2.3 AI for Efficiency and Risk Categorisation: Use of Classification in Resource Management

Classification techniques of health care are primarily aimed at predicting clinical risks, which include identifying those at risk for sepsis, contamination-related drug reactions, or hospital-acquired infections. In administration, classification is essential in categorising patients according to high costs or resource requirements. By categorising patients into cost tiers (low, medium, high), healthcare organisations can distribute resources, forecast expenses, and customise care plans according to clinical risk evaluation. Empirical evidence shows the efficacy of classification models such as Logistic Regression, Support Vector Machines, and tree-based algorithms in improving patient stratification considerably. For example, [10] applied the model of classification that enabled them to classify emergency department patients based on the complexity and cost of treatment that was expected to be provided to them, which led to improved staff and patient flow. Besides, [11] used Random Forest models on insurance claims databases to cluster patients by chronic disease risk and the accompanying health expenditures. The use of these models enabled improved targeting of high-risk patients, supported timely interventions and ultimately improved efficiency in care delivery at the expense of avoiding avoidable hospitalisations. Despite such accomplishments, the field has continued to render classification as disjoint from regression-based financial modelling, thereby preventing such methods from being integrated into wider healthcare management processes.

2.4 Research Gaps Identified

While research has made significant progress in regression models for estimating cost and classification models for risk stratification, the formulation of joint approaches encompassed under a single cohesive structure is yet to be done to a large extent. This underscores a critical deficiency, as the use of a dual task model would enable not only accurate expenditure estimation but also effective segmentation of patients based on risk levels [12]. The failure of integrated frameworks to have gaps has led to operational inefficiencies, which have led to fragmented decision-making strategies. In addition, the

10.48047/jocaaa.2023.31.04.26

literature tends to favour single metrics such as accuracy or R^2 instead of holistic model evaluation, recommended by complementary measures such as precision and recall and the F1-score. Consequently, models will likely lose important interpretability and generalisation capabilities, especially if working with class imbalance or skewed data distributions typical in healthcare settings. Engaging visualisation diagnostic aids, such as confusion matrices, ROC curves, and predicted-vs-actual plots, while promising for transparency and stakeholder involvement, are frequently overlooked in the analysis. One significant shortcoming is the inability to explain how models predict. Although the predictive capabilities are improved, there are limitations in clinical validation and approval from a regulatory point of view for Random Forest and XGBoost models because of the inherent nature of their black-box approach. More than a few studies do not employ explainability techniques such as SHAP or LIME to justify the basis for model predictions for clinical and administrative stakeholders [13]. Therefore, future studies must combine regression and classification while incorporating robust evaluation procedures and enhancing the interpretability of models.

3 Methodology

3.1 Proposed Framework

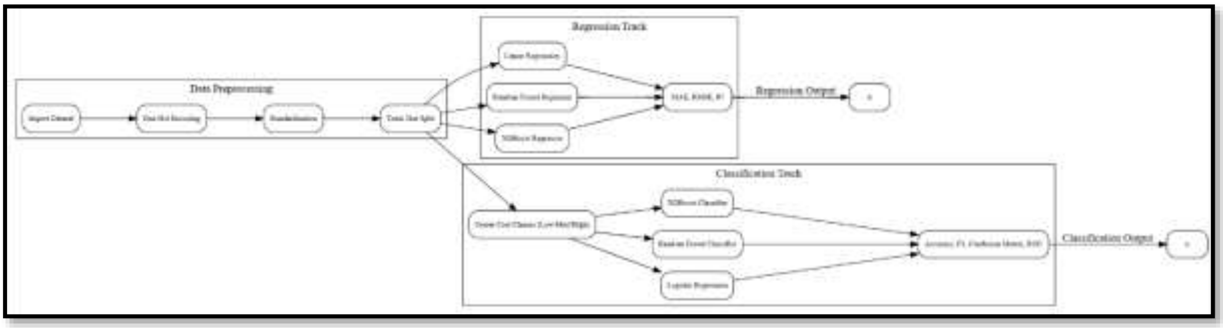

Figure 1: Proposed Methodology Diagram

Figure 1 Indicates a dual-track machine learning framework for healthcare cost and efficiency modelling. Data preprocessing begins with obtaining the dataset, one-hot encoding for characteristic variables, normalisation, and dividing it into training and testing. The next two independent workflows are created to be executed in parallel: The regression track uses Linear Regression, Random Forest, and XGBoost for the prediction of continuous medical charges and the classification track uses Logistic Regression, Random Forest, and XGBoost classifiers for assigning patients to cost-risk tiers (Low, Medium, High). Both tracks generate outputs scored in MAE/RMSE/R^2 for regression and accuracy, F1-score, confusion matrix, and ROC for classification to assist operational decision needs.

3.2 Dataset Description

This study utilised Medical Cost Personal Datasets comprising 1,338 individual records. Patient details regarding demographics, lifestyle aspects, and associated medical insurance payments are integrated for each record in the dataset. Its diagrammatic nature and extensive, cost-determining nature make this dataset prevalent in healthcare machine learning research. The dataset reflects core attributes including the age, sex, body mass index (BMI), no of dependent children, smokers and geographic region. These characteristics emerge repeatedly as dominant factors that explain changes in healthcare costing, and they are appropriate for regression and classification.

The dataset contains one primary continuous variable, charges, revealing an individual patient's overall medical insurance cost. The target charges were incorporated into regression models to calculate each patient's financial burden of care. charge_class was constructed as a classification-ready variable by aggregating the charges into three groups, namely Low, Medium, and High, for classification purposes.

10.48047/jocaaa.2023.31.04.26

The application of binning led to separate cost-efficiency brackets for patients whose efforts supported the multiclass classification. Both types of target variables in one dataset allow for assessing AI models in a predictive and stratification context.

3.3 Data Preprocessing

Several pre-processing techniques of data were applied before actual data processing to ensure the model is ready to be transformed into modelling and enhance overall model accuracy. At the onset, the categorical variables, namely, sex, smoker, and region, were each converted to one-hot encoding, which permitted them to be converted to binary numerical columns that could efficiently be used in ML algorithms. This strategy preserved the initial information displayed in categorical data without attributing false ordering, which might tarnish the model’s performance.

Feature scaling was then carried out by using scikit-learn's StandardScaler. This step was necessary because several machine learning techniques, including gradient-based and distance computation methods, are sensitive to differences in scale across features. Age and BMI were scaled to a standardised scale with a mean of 0 and a standard deviation of 1 to facilitate consistent learning, which benefited the performance of various algorithms.

Two separate data subsets were formed by dividing them along the axis of the target variable. The data was split into two subsets: one for regression tasks, where continuous charges were predicted, and the other for classification tasks that looked at charge_class. The regression and classification datasets were partitioned according to an 80-20 split into training and test segments using a fixed random seed for reproducibility. LabelEncoder was implemented on the target labels for the classification problem, allowing the use of multiclass classifiers such as Random Forest and XGBoost. Through such preprocessing remedies, data uniformity was retained, positively influencing the possibility of obtaining accurate and interpretable results from the models.

3.4 Regression Modelling Techniques

Three different regression models were selected to predict medical charges, from simple linear methods to intricate ensemble techniques that apply to predicting medical charges. Linear Regression, the first model used, is classed as a control to compare the performance of more sophisticated models. Keeping simplicity and interpretability, linear regression is still prone to linear connections, multicollinearity and heteroscedasticity, problems common in healthcare research.

The second model is called Random Forest Regressor, an ensemble learning, meaning that many decision trees are trained and their predictions are averaged at the end. It is appropriate for non-linear data, does not overfit, and is highly accurate, even though there are fewer training examples. Its inherent feature ranking also facilitates explanation because the most influential predictors will be emphasised.

The third model, XGBoost Regressor, uses gradient boosting to refine model errors iteratively, thus providing better accuracy in structured data handling. The model does not respond to outliers, is good at representing feature interactions, and has inherent regularisation, allowing control overfitting.

To ensure constant representation across models, each was trained using the same subset of data. Model performance was compared by using three established regression metrics:

- **Mean Absolute Error (MAE)**, with a mean estimate of the difference between predicted and actual charges.
- **Root Mean Square Error (RMSE)**, which intensifies the penalty for larger errors in predictions, and
- **R-squared (R^2)**, whose values show the percentage of variability in the target accounted for by the model.

10.48047/jocaaa.2023.31.04.26

The feature importance analysis for tree-based models was then performed alongside evaluating model performance to identify desirable smoking status, BMI, or age attributes in influencing medical cost forecasts. In such scenarios, interpretability is critical, particularly since stakeholders demand detailed explanations for existing predictions.

3.5 Classification Modelling Techniques

For categorising efficiency, a target variable called 'charge_class' was derived and divided into three cost categories: Low, Medium, and High. For this purpose, Quantile binning was used to define such cost tiers while ensuring that each class has approximately the same number of measurements, thereby reducing any possible model bias due to class imbalance.

Three separate models were used to address this challenge. Logistic Regression is the first classifier, which is linear, and there is no dependency of features here, and a logistic function is also used for determining the boundary between classes. Logistic regression is easy; therefore, it beats simpler models in first predictions and maintains interpretability due to the linear coefficients.

By adapting the same ensemble tree approach from regression, the second classifier, Random Forest, can be reconfigured for effective multiclass classification. Random Forest Classifier can easily take care of outliers and avoid over-fitting by using boot-strapping, and also provides for strong class differentiation through majority voting for the ensemble.

XGBoost Classifier is the equivalent applied in classification that goes hand in hand with the regression model we reviewed above. The classifier is based on gradient-boosted decision trees with an optimised training speed and a combination of model accuracy. XGBoost Classifier is known for its success in Kaggle competitions and the healthcare environment, where it effectively handles complex feature relations as well as the state of multiclass issues.

Model evaluation included several key metrics:

- **Accuracy**, the overall proportion of correct predictions,
- **Precision** (macro-averaged), the ability of the classifier to avoid false positives
- Recall (macro), which measures the model's sensitivity for each class's true positives, and
- **F1-score (macro)**, based on the harmonic mean of precision and recall, reflecting the balanced model's performance.

Misclassification Tendencies were discovered by examining Confusion matrices and Receiver Operating Characteristic (ROC) curves conducted using a one-vs-rest strategy to analyse the discriminative ability of the models across classes. The Area Under the Curve (AUC) provided an accurate picture of model performance, particularly when facing issues like imbalanced classes or messy signals.

3.6 Tools and Platforms Used

The entire modelling, analysis, and deployment happened in Google Colab, a cloud-based Jupyter notebook service that enables smooth code deployment and sharing. High-performance Python libraries were at our disposal, and GPU acceleration was available whenever needed. The project used Python 3.10 and a series of open-source libraries developed for machine learning and data visualisation tasks.

Data manipulation and preprocessing were performed using Pandas and NumPy, and visualisations were carried out with Matplotlib and Seaborn. Popular algorithms such as Logistic Regression and Random Forest were applied using scikit-learn, whilst gradient boosting models were implemented using XGBoost. In addition, the LabelEncoder, StandardScaler, and essential evaluation metrics like classification_report and roc_auc_score were included to guarantee proper pipeline development and

10.48047/jocaaa.2023.31.04.26

evaluation. Using low-threshold resources and making code reproducible, the team could examine model success and gain insights that would apply to healthcare.

4 Result and Discussion

4.1 Exploratory Data Analysis

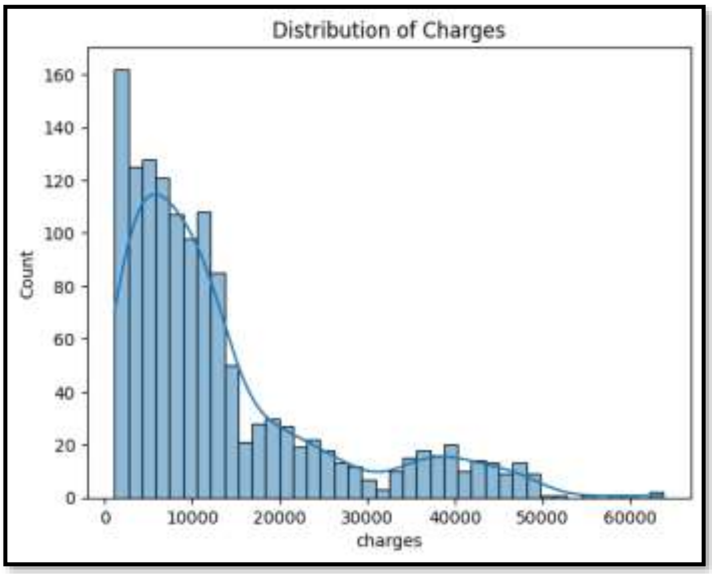

Figure 2: Distribution of Charges

From Figure 2 It can be seen that healthcare charges are right-skewed- most patients paid less than $15000, but others faced much higher charges. The skewed nature of the observations sends a signal towards the presence of extreme cost outliers, which can be the cause of chronic health problems or habits such as smoking. KDE curve shows that many costs are concentrated in the $5000-$10,000 range, aligning with standard outpatient and preventive services. The character of the skew shows that stratified prediction models are necessary.

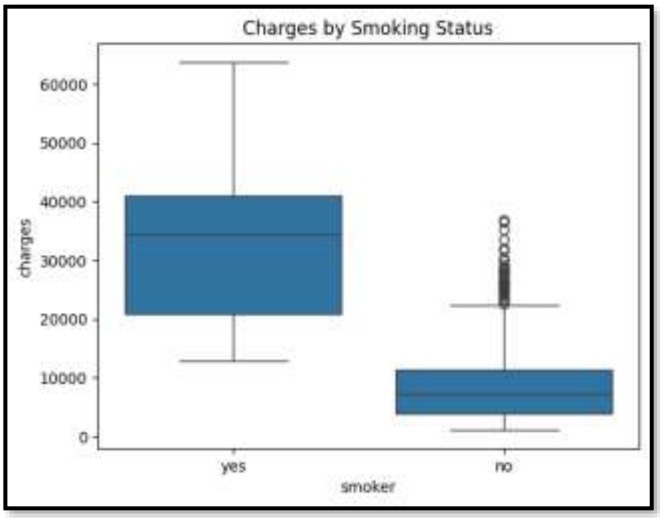

Figure 3: Charges by Smoking Status

10.48047/jocaaa.2023.31.04.26

Figure 3 shows a huge financial deficit in healthcare costs between smokers and non-smokers. Median and IQR values for smokers are significantly higher, and their median cost is far beyond $30,000, while for non-smokers, the median is still below $10,000. The more variation and clustering of outliers in the smokers shows that many are experiencing a succession of costly health issues that are often associated with comorbidities, e.g., cardiovascular or respiratory conditions. This shows the importance of smoking status in both estimating healthcare costs and results for managing healthcare resources.

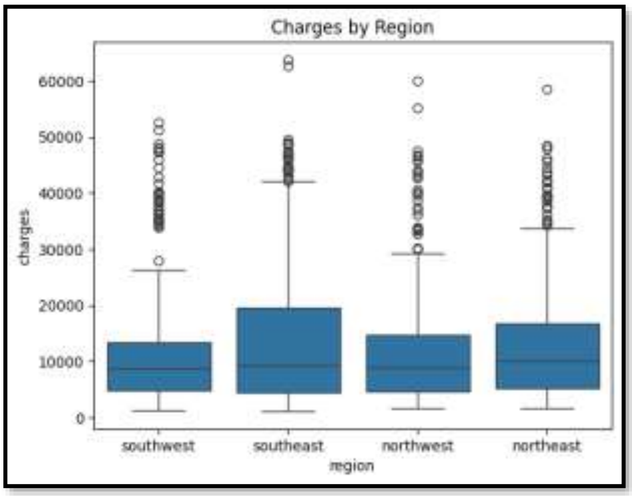

Figure 4: Charges by Region

Figure 4 provides a comparison of medical charges within four U.S regions. Whilst the median costs per region are similar, the upper quartile is slightly higher for the southeast, which is associated with more variance of costs. There are outliers in every region, but the extreme maximum charges are most evident in the Southeast and Northeast regions. It may result from differences in the healthcare pricing systems, regional health profiles or healthcare policies between the areas. From these results, 'region' deserves inclusion as a categorical predictor in cost and classification models.

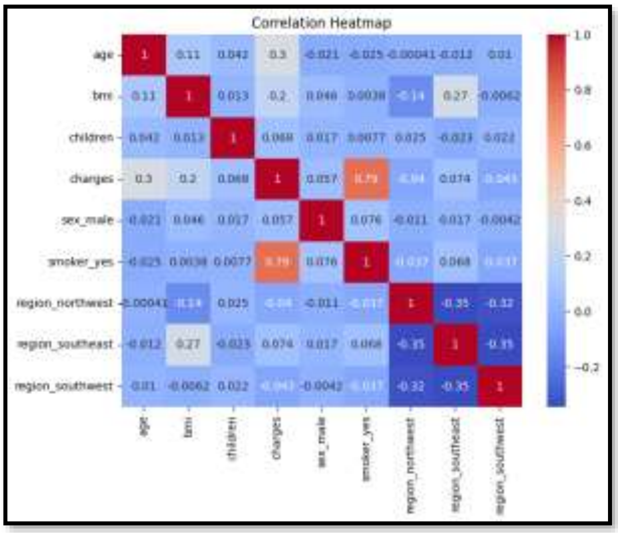

Figure 5: Correlation Heatmap

10.48047/jocaaa.2023.31.04.26

Figure 5 shows the pairwise Pearson correlation coefficients of each encoded variable. As the data indicates, 'charges' positively correlate to 'age' (0.30) and 'bmi'. The lack of or inverse correlations in region encodings reveal that single regions do not significantly impact costs individually, although a synergistic interference is possible. Judging from these findings, it is reasonable to name smoking status and BMI as the first predictors in regression and classification models.

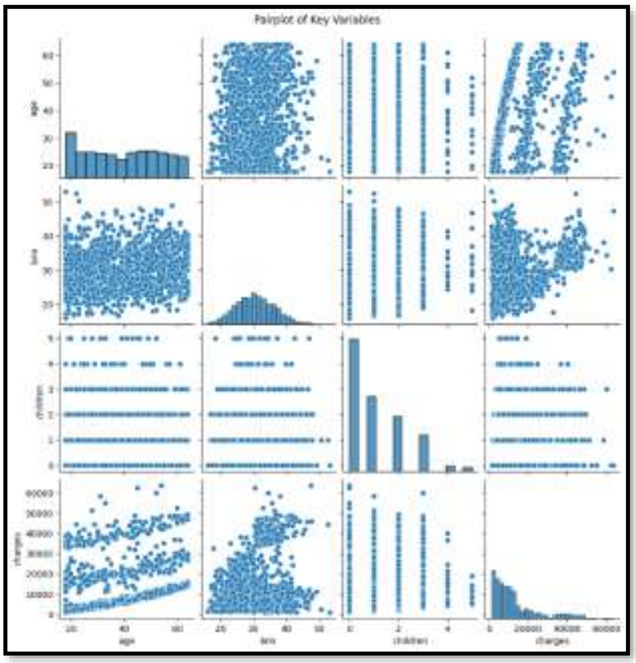

Figure 6: Pairplot

Figure 6 provides detailed insight into the interrelations of age, BMI, children and charges. The graph shows that charges have increased systematically with age and BMI, leading to an upward movement. Instead of age, the BMI distributions are bell-shaped, as opposed to age, which predominantly favours younger people. There is little clear relationship between children and charges, indicating a compromised predictive power. The monetary spread of data points suggests that using nonlinear algorithms, such as Random Forest and XGBoost, is relatively efficient.

4.2 Model Evaluations

10.48047/jocaaa.2023.31.04.26

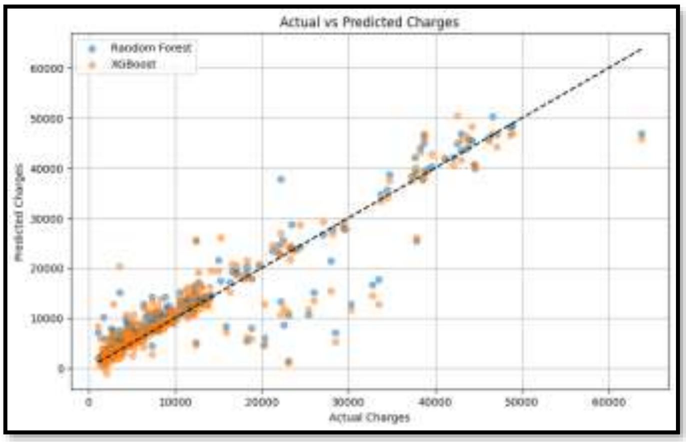

Figure 7: Actual vs Predicted Charges

In Figure 7, a graph has been used to compare the actual versus predicted charges from the Random Forest and XGBoost models. Both models ' high alignment with the 45° reference line validates their overall prediction accuracy. The small underestimation is seen at the upper bounds of the charge predictions, displaying the difficulty in accurately predicting such values. Random Forest is more scattered in the mid-range, but XGBoost clusters well together. This implies that under segments that exhibit higher variability, XGBoost provides better results.

```
--- Linear Regression ---
MAE: 4181.19
RMSE: 5796.28
R²: 0.78

--- Random Forest ---
MAE: 2545.21
RMSE: 4567.96
R²: 0.87

--- XGBoost ---
MAE: 2765.75
RMSE: 4840.94
R²: 0.85
```

Figure 8: Model Metrics

Figure 8 shows the performance results of the three regression models when making predictions. Linear Regression achieved an R^2 of 0.78 and had the most significant errors, proving it was challenging to deal with non-linear data. Relative to this, Random Forest presents better results with an R^2 of 0.87 and an RMSE of ~4568; XGBoost registers an R^2 of 0.85 and an RMSE of ~4841. Such results reveal how tree-based ensemble models are superior in predicting feature interactions and correcting skewed distributions for healthcare costs.

10.48047/jocaaa.2023.31.04.26

Logistic Regression:

	precision	recall	f1-score	support
0	1.00	0.64	0.78	84
1	0.92	0.91	0.92	105
2	0.70	0.97	0.81	79
accuracy			0.85	268
macro avg	0.87	0.84	0.84	268
weighted avg	0.88	0.85	0.85	268

Figure 9: Logistic Regression Classification Report

The performance of Logistic Regression is reported to be this accurate, with a value of 85%, as per the classification metrics. Class 1 (medium cost) has a high precision (0.92), recall (0.91), whereas class 0 (low cost) has perfect precision (1.00), but low recall (0.64), indicating overestimation. In Class 2 (high cost), what is high recall (0.97) is the opposite of low precision (0.70), suggesting class imbalance or overlapping feature boundaries (Figure 9). The macro average F1-score of 0.84 shows well-balanced efficiency of classes, but uneven efficiency also exists.

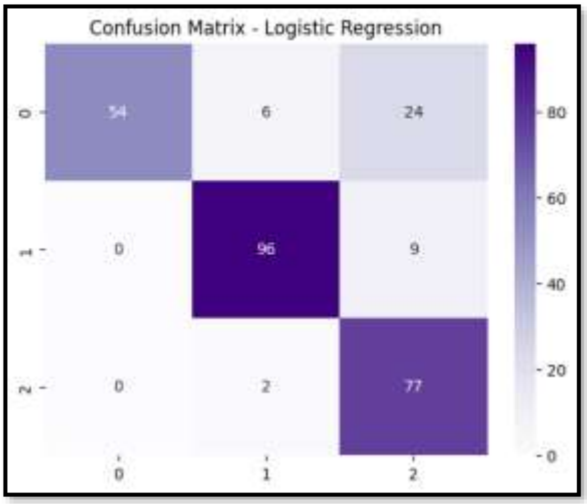

Figure 10: Confusion Matrix – Logistic Regression

From the logistic regression confusion matrix, there is an offsetting emphasis on distinct divisions, particularly between medium (class 1) and high (class 2) charges. Class 1 shows 96 accurate predictions as well as minimal errors (Figure 10). However, the model classified 24 cases from class 0 as class 2, which could mean an overly optimistic approach to cost severity. The matrix nevertheless indicates that the model generalises adequately between categories, particularly in areas with clear boundaries between the medium and high charge levels.

Random Forest:

	precision	recall	f1-score	support
0	0.94	0.81	0.87	84
1	0.92	0.95	0.93	105
2	0.84	0.92	0.88	79
accuracy			0.90	268
macro avg	0.90	0.90	0.90	268
weighted avg	0.90	0.90	0.90	268

10.48047/jocaaa.2023.31.04.26

Figure 11: Random Forest Classification Report

Random Forest provides classification accuracy of 90%, showing similar precision, recall and F1 score value for each class. The model is virtually perfect for cases in the medium-cost category, which fall into Class 1, with an F1 score of 0.93. The high-cost class (Class 2) also tends to have a good F1-score of 0.88 (Figure 11). A macro average and a weighted average both achieve 0.90, highlighting the model's strength in multiclass charge stratification and preparation for adoption in healthcare triage systems.

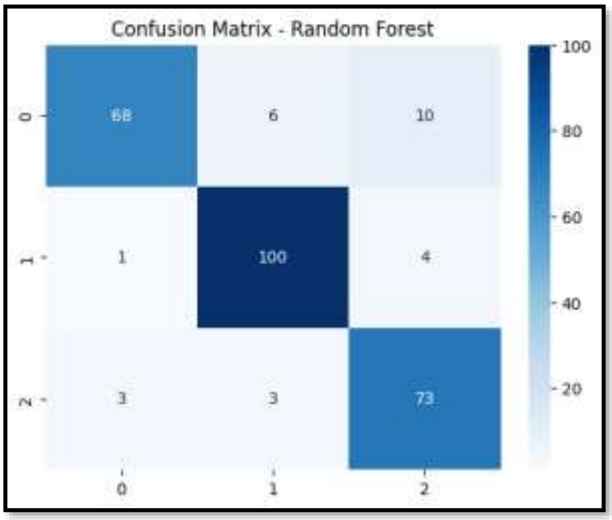

Figure 12: Confusion Matrix – Random Forest

The Random Forest confusion matrix validates the close performance of the model in discerning all three levels of medical charges. For Class 0 (low cost) and Class 2 (high cost), the model produces few false positives and negatives, with most observations along the diagonal. In Class 1, there is almost no error anywhere between precision and recall, and some areas of misclassification occur (Figure 12). This evidence supports Random Forest's better stability because Logistic Regression has higher error rates in identifying charges, which are close in value.

XGBoost:

	precision	recall	f1-score	support
0	0.91	0.82	0.86	84
1	0.92	0.95	0.93	105
2	0.84	0.89	0.86	79
accuracy			0.89	268
macro avg	0.89	0.89	0.89	268
weighted avg	0.89	0.89	0.89	268

Figure 13: XGBoost Classification Report

XGBoost attains a high classification efficiency, as high as 89%, similar to Random Forest. The model again exhibits remarkable precision (0.92) and recall (0.95) for Category 1 (medium costs). The class 2 F1 score is 0.86, lower than the Random Forest yields. The macro and weighted averages on all metrics are 0.89, and there is consistent and reliable generalisation (Figure 13). In applications where speed of training and scalability are interesting, XGBoost presents a highly suitable alternative.

10.48047/jocaaa.2023.31.04.26

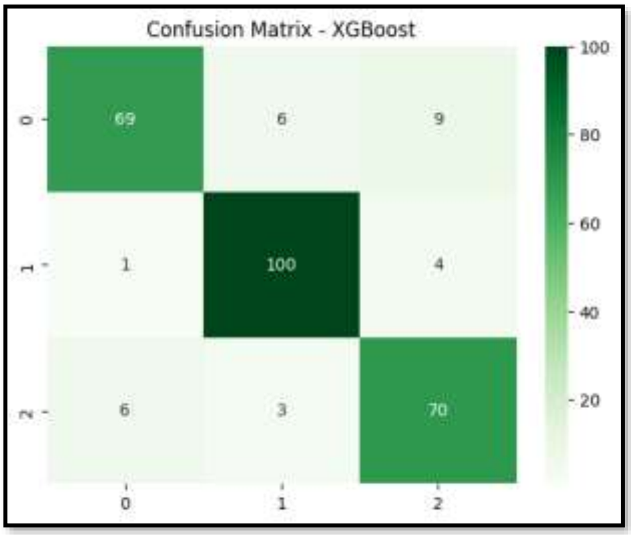

Figure 14: Confusion Matrix – XGBoost

XGBoost accurately identifies most class 1 instances—only five out of 105 were incorrect. Compared to Random Forest, Class 0 and Class 2 exhibit slightly higher misclassification rates, producing nine incorrect predictions for low-cost and 6 for high-cost instances (Figure 14). Nevertheless, predictions remain mostly aligned along the diagonal. It proves that XGBoost effectively segments cost categories, despite data imbalance and complex healthcare cost patterns.

4.3 Model Comparison

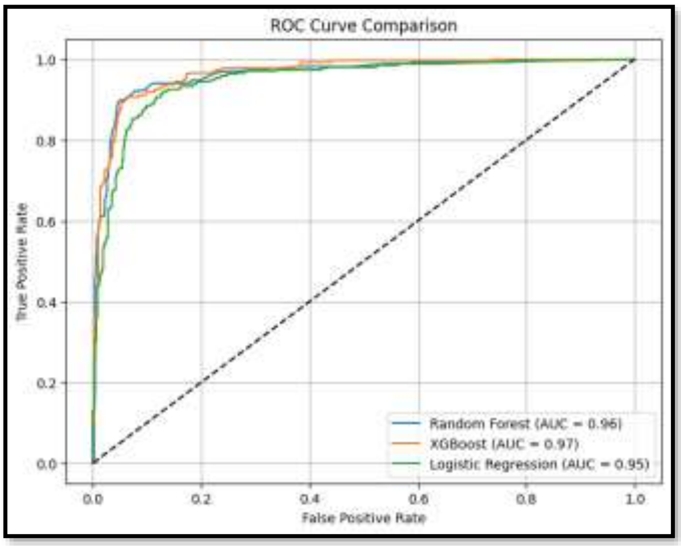

Figure 15: ROC Curve Comparison

An ROC curve comparison is presented to compare how each model classifies data in a binary fashion using all the classes as separate reference categories. XGBoost outperforms the other models with the most considerable AUC value (0.97), showing better separability between cost classes. With AUC figures of 0.96 and 0.95, Random Forest is the second-best, followed by Logistic Regression (Figure 15). While all models show high levels of discrimination, the curve shows that XGBoost is superior in sensitivity and specificity. This result corroborates XGBoost's usefulness in situations where class prediction errors might substantially impact treatment decisions.

10.48047/jocaaa.2023.31.04.26

🔍 Classification Performance Comparison (Macro Average):

	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest	0.8993	0.9003	0.8953	0.8953
1	XGBoost	0.8918	0.8896	0.8866	0.8871
2	Logistic Regression	0.8470	0.8744	0.8439	0.8387

Figure 16: Combined Classification Performance

The performance table briefly overviews the three models' performance, using macro-averaged metric values. Among the examined approaches, Random Forest yields the highest accuracy at 0.8993 and F1 score at 0.8953, with XGBoost showing near parity. Although Logistic Regression is more efficient due to its simplicity, it did not perform as well as the other two models in any aspect, as demonstrated in Figure 16. This table indicates that Random Forest is best suited for classification problems that require accuracy and interpretability, specifically where class distinctions may be unclear, as happens in healthcare data.

4.4 Discussion

The results of this study suggest that AI-driven predictive analytics is a powerful approach to handling two key issues in healthcare. predicting how much each person will spend on medical care and assigning patients to cost-efficiency categories. A structured modelling pipeline used on actual healthcare insurance data shows that Random Forest and XGBoost, as ensemble machine learning methods, perform markedly better than conventional linear models in both mathematical modelling and classification tasks. The regression models devised to predict real healthcare charges found that Random Forest Regressor performed best, achieving both the highest R^2 of 0.87 and the lowest error measurements across MAE and RMSE, reflecting its strength in handling intricate, nonlinear linkages between variables, including age, BMI, and smoking status. The findings support the contention that healthcare cost prediction must transition away from relying exclusively on simple linear models or rule-based approaches. As a result, data-centred, flexible techniques are indispensable for reliably forecasting individual financial risk.

The classification part of the study assigned patients into low, medium, or high-cost categories, depending on their predicted charges. This section shows that Random Forest produced the most outstanding results, reaching an accuracy of 89.93% and an F1-score of nearly 0.90 on a macro scale. Although Logistic Regression is simpler and more understandable, it lacks precision in recall, especially for patients with lower predicted costs. ROC-AUC scores demonstrated the ensemble models' generalizability across cost categories, and XGBoost demonstrated a slightly better capacity to separate among classes. The capacity of these classifiers is especially relevant in medical care, as it allows patients to be sorted according to their use of resources or expected costs.

The importance of these results surpasses just determining how accurate the models are. When such AI models are integrated into major healthcare infrastructure, like information systems, insurance networks, or national health planning units, they could empower earlier and more effective decision-making. Predictive regression models enable financial planners to predict future financial demands, detect people at high risk of incurring significant costs, and streamline insurance premium structuring. Moreover, classification technology can be applied directly to triage, so health professionals can better allocate resources to patients needing cost containment. Applying this model in a hospital, patients with a high predicted cost may be identified and placed into multidisciplinary care pathways in advance, supporting better results and decreasing long-term costs.

10.48047/jocaaa.2023.31.04.26

In addition, policy-minded applications of these models can help advance population-level analytics, revealing inefficiencies related to demographic characteristics, regions, or behavioural risks. The clear link between smoking status and BMI, and future health costs, underlines the necessity of risk reduction measures and behaviour interventions. Subsequently, this information can direct specific health promotion strategies or regional measures to alleviate avoidable expenditures on high-cost healthcare. In addition, these models' interpretable and tool-based construction means they can be utilised at scale and reliably applied to significant real-time sources, such as electronic health records or national claims banks. The results demonstrate that AI can reliably support cost and efficiency modelling and indicate clear steps towards translating these approaches into clinical practice.

5 Conclusion

This study evaluated how well AI-driven predictive analytics can address two main healthcare challenges. forecasting costs and assigning patients to provider need categories with high precision and efficiency. This research, by utilising a dual-modelling structure for regression and classification, clearly shows that ensemble learning methods like Random Forest and XGBoost achieve much better results in prediction and stratification than traditional linear methods. Reliance on a widely available insurance dataset, which contains both demographic and lifestyle features, made it possible to provide a hands-on demonstration of extracting meaningful insights from easily interpretable variables.

In the regression analysis, Random Forest delivered the top performance, reflected in an R^2 score of 0.87 and the best results for MAE and RMSE. This finding reveals that Random Forest stands out by accurately capturing non-linear relationships and managing feature interactions, attributes widespread in healthcare datasets. Compared to Random Forest, XGBoost matched performance overall, but had somewhat lower consistency in identifying outliers. Although linear regression is frequently used as a starting point, it has difficulty capturing the skewed pattern of healthcare charges and does not give sufficient attention to higher-cost cases. These results emphasise that conventional statistical models are insufficient for understanding healthcare commodity patterns.

Patients were grouped into low, medium, and high-cost categories, based on predictions of their charges, within the framework's classification component. Random Forest was once more the strongest performer, achieving an accuracy of 89.93% and an F1-score of 0.8953 when measured according to macro-averaging. Despite being easy to interpret, Logistic Regression tends to misclassify those with low costs, which may not be suitable for clinical triage. XGBoost achieved both interpretable results and strong performance, as confirmed by the highest AUC (0.97) achieved in ROC analysis. The reported results significantly reinforce that ensemble models deliver accurate predictions and the robustness and interpretability needed for practical clinical use.

In a broader view, this work shows the potential applications of machine learning approaches within healthcare systems. With forecasted charges, hospitals and insurers can actively plan their resource use, create customised programs to manage costs, and price premiums according to predicted risks. When patients are assigned to cost-efficiency groups, healthcare providers can better target treatments for them, work with multidisciplinary teams for those with high costs, and simplify workflow processes. Enhancing efficiency and outcomes for patients confronts two main healthcare challenges at once.

Ultimately, the research presented proposes a scalable and interpretable approach, enabling replication to aid in planning and operational choices in healthcare environments. The framework validates that AI can contribute to clinical outcome prediction and financial management in healthcare. As resource constraints grow in healthcare, this dual-track AI model is essential in matching predictive analytics to ongoing system requirements.

6 Future Work

This paper recommends that healthcare organisations, regardless of their type, including hospitals, insurers, or public health agencies, use ensemble-based AI solutions like Random Forest and XGBoost for financial planning and patient assignment work. Their performance is reliable, requires only slight

10.48047/jocaaa.2023.31.04.26

parameter optimisation, and can be implemented using Python tools. These models can be used in current clinical decision support tools or insurance claims frameworks for improved prediction of high-cost patients and more effective triage.

Further work may be pursued along several valuable directions. First, model performance and relevance to clinical settings could improve if we extend the input features to clinically relevant variables, such as diagnosis codes, medication lists, and lab test results. Second, using time-series models, including LSTM and GRU, could make forecasting ongoing or repeated charges possible, which is valuable for helping with chronic conditions. A third priority is to incorporate explainability into our models, such as by using SHAP or LIME, to ensure clear insight, particularly in circumstances where interpretability is essential. Testing the framework on extensive, actual EHR sets, for example, MIMIC-IV, would reinforce its generalizability and make it more practically useful.

The core contribution of this paper lies in its demonstration that a unified AI framework can simultaneously predict individual patient costs and tier patients based on their efficiency. Unlike most earlier approaches that focused on only one way to model healthcare costs, this study combines two models to address financial and operational challenges. It proves that machine learning boosts prediction performance and concurrently generates insights that can guide resource assignment, budget setting, and clinical preparation. As a result, the study connects research in AI with the practical challenges faced in health care transformation.

References

1. Organization, W.H., *Global spending on health: rising to the pandemic's challenges*. 2022: World Health Organization.
2. Huang, Y., et al., *Application of machine learning in predicting hospital readmissions: a scoping review of the literature.* BMC medical research methodology, 2021. **21**: p. 1-14.
3. Vimont, A., H. Leleu, and I. Durand-Zaleski, *Machine learning versus regression modelling in predicting individual healthcare costs from a representative sample of the nationwide claims database in France.* The European Journal of Health Economics, 2022. **23**(2): p. 211-223.
4. Knevel, R. and K.P. Liao, *From real-world electronic health record data to real-world results using artificial intelligence.* Annals of the Rheumatic Diseases, 2023. **82**(3): p. 306-311.
5. Xie, F., et al., *Deep learning for temporal data representation in electronic health records: A systematic review of challenges and methodologies.* Journal of biomedical informatics, 2022. **126**: p. 103980.
6. De Panfilis, L., et al., *AI-based clinical decision-making systems in palliative medicine: ethical challenges.* BMJ Supportive & Palliative Care, 2023. **13**(2): p. 183-189.
7. Langenberger, B., T. Schulte, and O. Groene, *The application of machine learning to predict high-cost patients: A performance-comparison of different models using healthcare claims data.* PloS one, 2023. **18**(1): p. e0279540.
8. Bharadwaj, H.K., et al., *A review on the role of machine learning in enabling IoT based healthcare applications.* IEEE Access, 2021. **9**: p. 38859-38890.
9. Verma, M., J. Singh, and S. Kumari. *Enhancing heart disease prediction with ensemble deep learning and feature fusion in a smart healthcare monitoring system*. in *International Conference on Advanced Computing and Intelligent Technologies*. 2023. Springer.
10. Zaerpour, F., et al., *Patient classification based on volume and case-mix in the emergency department and their association with performance.* Health care management science, 2020. **23**: p. 387-400.
11. Nicolet, A., et al., *Exploring patient multimorbidity and complexity using health insurance claims data: a cluster analysis approach.* JMIR Medical Informatics, 2022. **10**(4): p. e34274.
12. Amiri, Z., et al., *The personal health applications of machine learning techniques in the internet of behaviors.* Sustainability, 2023. **15**(16): p. 12406.
13. Rane, N., S. Choudhary, and J. Rane, *Explainable artificial intelligence (XAI) in healthcare: Interpretable models for clinical decision support.* Available at SSRN 4637897, 2023.

10.48047/jocaaa.2023.31.04.26