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

# Patient Expenditure Prediction in Healthcare Sectors

**Subhani Shaik[1], P. Vinay Kumar[2], Rohith Viswanathan[3], M. Ganesh[4]**

[1]Associate Professor, Department of IT, Sreenidhi Institute of Science & Technology (A), Hyderabad, India

[2]IV B. Tech Student, Department of IT, Sreenidhi Institute of Science & Technology (A), Hyderabad, India
E-mail: *subhanicse[at]gmail.com*

**Abstract:** ***Predicting the expenses for patient expenditure in health sectors became an important task with various applications such as provider profiling, accountable management, and medical payment Adjustment. Previous approaches mainly deals with manually designed features and linear regression-based models, which require massive medical domain knowledge and show limited predictive performance. This paper gives us a multi-view deep learning framework which can help to predict upcoming healthcare expenses at the individual level which are based on historical data. Our multi-view approach can accurately model the mixed information, including patient demographic features, medical codes, drug usages, and facility utilization. We conducted forecasting of expense tasks on a day-to-day pediatric dataset that contains more than 390,000 patients. The empirical results displays that our proposed method outperforms all baselines for medical expenses calculation. These findings help toward better preventive care and accountable care in the healthcare domain.***

**Keywords:** Administrative claims data, deep learning, electronic health records, expenditure evaluation, machine learning

## 1. Introduction

The rise of healthcare sector expenses constitutes a challenging approach to health management and healthcare associations. Since it was clarified by the CMS Services of Medicine, the United States for the national health expenditure (NHE) grew4600 to approx $3600 billion in 2018( i.e., One $ per person) and established under the calculation for 18 of GDP. Specifically, Most of the services spending grew64 to$7502 million, and CMS service grew by30 to$5974 million. The health management system is mostly about to come inequitable unless medical price growing scale is kept under the observation. (1). It is prominent to take utmost observation of the health management expenditure rise and minimizing the cost of medicine for people. An Administrative data, a variety kind of EHR records authenticated for bills record purposes, maintains the length wise patient records including structure of people, judgments, procedures, specifics, installation National Health. Expend Data affirms the data as one of the accurate available sources for estimating cases of individual conditions. Here an adding amount of affirmed data shows the latest, accurate approach to attack management expense issues. Using the actual claims, we may able to produce collection of data models to reveal important perceptivity which are taken part through the pattern of expenses. Particularly, the price of exact medicines prophetic model in an individual position might take into consideration of the cases with huge threat also it gives an awesome quality of supervision. To the various initiations for expenses on patients based vaticination generally calculate on manmade substances and direct retrogression- grounded features (2), (3). In this type of case, the Direct Cost Group which in short DCG(4) feature applies direct retrogression in order to prognosticate expense on health grounded through personal orders which are managed by sphere expertise people. (5) Using CART grounded on the collections of medicinal principles and manmade cost features was

## 2. Literature Survey

Y. Zhao et al., "Predicting pharmacy costs and other medical costs using diagnoses and drug claims," Med. Care, vol. 43, pp. 34–43, 2005 During congressional debate over the Medicare Part D prescription drug benefit, much attention was focused on nominal benefit design. Relatively little attention was paid to details about how plans would operate, such as the design of drug formularies. Yet, formularies will be important tools for controlling costs, and may be as important as nominal benefit design in determining enrollees' access to medications and out-of-pocket costs. We describe Part D plan incentives and how they may influence formulary design, and then provide recommendations for Part D formulary implementation. We encourage the Centers for Medicare & Medicaid Services (CMS) to develop standardized tools to provide physicians and patients with up-to-date and easily accessible information about covered drugs on each plan's formulary (perhaps via a central website) and a national set of easy-to-follow procedures for reconsideration and appeals. Such efforts should reduce administrative burden and better allow physicians to help patients obtain needed medications.

M. A. Morid, O. R. L. Sheng, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Healthcare cost prediction: Leveraging fine-grain temporal patterns," J. Biomed. Inform., vol. 91, 2019, Art. no. 103113 methods by using structural temporal pattern detection that captures global and local temporal trends and to demonstrate these improvements in the detection of acute kidney injury (AKI). Using the Medical Information Mart for Intensive Care dataset, containing 22,542 patients, we extracted both global and local trends using structural pattern detection methods to predict AKI (ie, binary prediction). Classifiers were built on 17 input features consisting of vital signs and laboratory test results using state-of-the-art models; the optimal classifier was selected for

comparisons with previous approaches. The classifier with structural pattern detection features was compared with two baseline classifiers that used different temporal feature extraction approaches commonly used in the literature: (1) symbolic temporal pattern detection, which is the most common approach for multivariate time series classification; and (2) the last recorded value before the prediction point, which is the most common approach to extract temporal data in the AKI prediction literature. Moreover, we assessed the individual contribution of global and local trends. Classifier performance was measured in terms of accuracy (primary outcome), area under the curve, and F-measure. For all experiments, we employed 20-fold cross-validation. Random forest was the best classifier using structural temporal pattern detection. The accuracy of the classifier with local and global trend features was significantly higher than that while using symbolic temporal pattern detection and the last recorded value (81.3% vs 70.6% vs 58.1%; *P*<.001). Excluding local or global features reduced the accuracy to 74.4% or 78.1%, respectively (*P*<.001). Classifiers using features obtained from structural temporal pattern detection significantly improved the prediction of AKI onset in ICU patients over two baselines based on common previous approaches. The proposed method is a generalizable approach to predict AEs in critical care that may be used to help clinicians intervene in a timely manner to prevent or mitigate AEs.

## 3. Problem Definition

Here, the given paper demonstrates on a purpose of deep literacy based on Multi View frame that can collect the miscellaneous information onto the electronic data. This frame consolidates various fields of data as different View. Notably, the proposed model grasp a FNN to bed the non-statistical features, an conscious- grounded bidirectional intermittent Neural N/W to capture the series installation functioning, and a stratified attention on linear network used for leveraging medicinal acquisition. So here scrutiny membrane allows providing the significance of giving input variables and demonstrates an elucidation of the forecast outraised.

## 4. Methodology

**Data Collection**
The data is extracted by the group of Medical team to respective administration office in the specified hospitals and gathering details of each and every bill the patient incurred and respective taxes and charges hospitals incurred in it. It is a historic data which are from 2017 to till now. This dataset consists of more than 8.5 Million medical records which are taken from 450 thousands of patients only from Jan 2015 to Dec 2017.

Since, due to the rules and policy from the National wide Health, the sources cannot be extracted fully as some of it can be very confidential.

**Data Preprocessing**
Data pre-processing is a just a other factor in the data mining process. The phrase "garbage in, garbage out" is the sentence which can be introduced to data mining and machine learning projects. Gathering of Data are often controlled weak, resulting outbound values, irrelevant data combinations, and missing values, etc.

**Training and Testing**
Training: The observations in the training set of a given Dataset form the experience that the algorithm uses to learn. In supervised type of Machine learning, several observation consists of an likely output variable and lot of observed input variables.

Testing: The test set is a useful observation that is used to evaluate the performance of the model using some metric. It is necessary that no observations are included in the test set which are introduced in the Training set.

This proposed method can be divided into following segments. These segments are, identify probability of patient having specific diseases, Data pre-processing, feature extraction, introducing machine learning algorithms, model efficiency evaluation, testing and deploying to a web application.

**Data Collection**
Data is collected from the history of medical reports which are available in Kaggle and organized in a well structured manner where the dataset can be understood by everyone. With the dataset which contains about 963 rows and 11 columns in which about approximately 10000 samples are available for pre-processing the data. Collecting the data that increases the data from the external resources by appending other data.

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Age | Diabetes | BloodPressureProblems | AnyTransplants | AnyChronicDiseases | Height | Weight | KnownAllergies | HistoryOfCancerInFamily | NumberOfMajorSurgeries | PremiumPrice |
| 2 | 45 | 0 | 0 | 0 | 0 | 155 | 57 | 0 | 0 | 0 | 25000 |
| 3 | 60 | 1 | 0 | 0 | 0 | 180 | 73 | 0 | 0 | 0 | 29000 |
| 4 | 36 | 1 | 1 | 0 | 0 | 158 | 59 | 0 | 0 | 1 | 23000 |
| 5 | 52 | 1 | 1 | 0 | 1 | 183 | 93 | 0 | 0 | 2 | 28000 |
| 6 | 38 | 0 | 0 | 0 | 1 | 166 | 88 | 0 | 0 | 1 | 23000 |
| 7 | 30 | 0 | 0 | 0 | 0 | 160 | 69 | 1 | 0 | 1 | 23000 |
| 8 | 33 | 0 | 0 | 0 | 0 | 150 | 54 | 0 | 0 | 0 | 21000 |
| 9 | 23 | 0 | 0 | 0 | 0 | 181 | 79 | 1 | 0 | 0 | 15000 |
| 10 | 48 | 1 | 0 | 0 | 0 | 169 | 74 | 1 | 0 | 0 | 23000 |
| 11 | 38 | 0 | 0 | 0 | 0 | 182 | 93 | 0 | 0 | 0 | 23000 |
| 12 | 60 | 0 | 1 | 0 | 0 | 175 | 74 | 0 | 0 | 2 | 28000 |
| 13 | 66 | 1 | 0 | 0 | 0 | 186 | 67 | 0 | 0 | 0 | 25000 |
| 14 | 24 | 0 | 0 | 0 | 0 | 178 | 57 | 1 | 0 | 1 | 15000 |
| 15 | 46 | 0 | 1 | 0 | 0 | 184 | 97 | 0 | 0 | 0 | 35000 |
| 16 | 18 | 0 | 0 | 1 | 0 | 150 | 76 | 0 | 0 | 1 | 15000 |
| 17 | 38 | 0 | 0 | 0 | 0 | 160 | 68 | 1 | 0 | 1 | 23000 |
| 18 | 42 | 0 | 0 | 0 | 1 | 149 | 67 | 0 | 0 | 0 | 30000 |
| 19 | 38 | 1 | 0 | 0 | 0 | 154 | 82 | 0 | 0 | 0 | 23000 |
| 20 | 57 | 1 | 0 | 0 | 0 | 156 | 61 | 0 | 0 | 0 | 25000 |
| 21 | 21 | 0 | 1 | 0 | 0 | 186 | 97 | 0 | 0 | 0 | 15000 |
| 22 | 49 | 1 | 0 | 0 | 0 | 160 | 97 | 0 | 0 | 2 | 28000 |
| 23 | 20 | 1 | 0 | 0 | 0 | 181 | 81 | 0 | 0 | 0 | 15000 |
| 24 | 35 | 0 | 0 | 0 | 0 | 163 | 92 | 0 | 0 | 1 | 32000 |
| 25 | 35 | 0 | 1 | 0 | 0 | 175 | 83 | 0 | 0 | 1 | 23000 |



**Figure 1:** DATASET of Medical Expenses

Describing the attributes about its nature where the dataset generally deals with the accurate data for the better preprocessing purpose. This data which are able to use for digital marketing based usage that can be filtered are varied on several regions.

**Data representation**

Data can be represented in various aspects as the dataset contains a collection of data having more than 950 rows and 11 columns that can be learned to provide an insight of the data in terms of Visualization and minor modifications of the data. Some of the changes that can be made based on various aspects that are Categorical data, Graphical Data, Numeric Data and Binary data which comes under the Primitive Feature Vector.

**Data Visualization**

Visualization of the data can be done by using seaborn library. So, Here the data visualizes that either the patient undergone in any of the criteria or not that is based on the categorical conversion where True is 1 and False is 0.

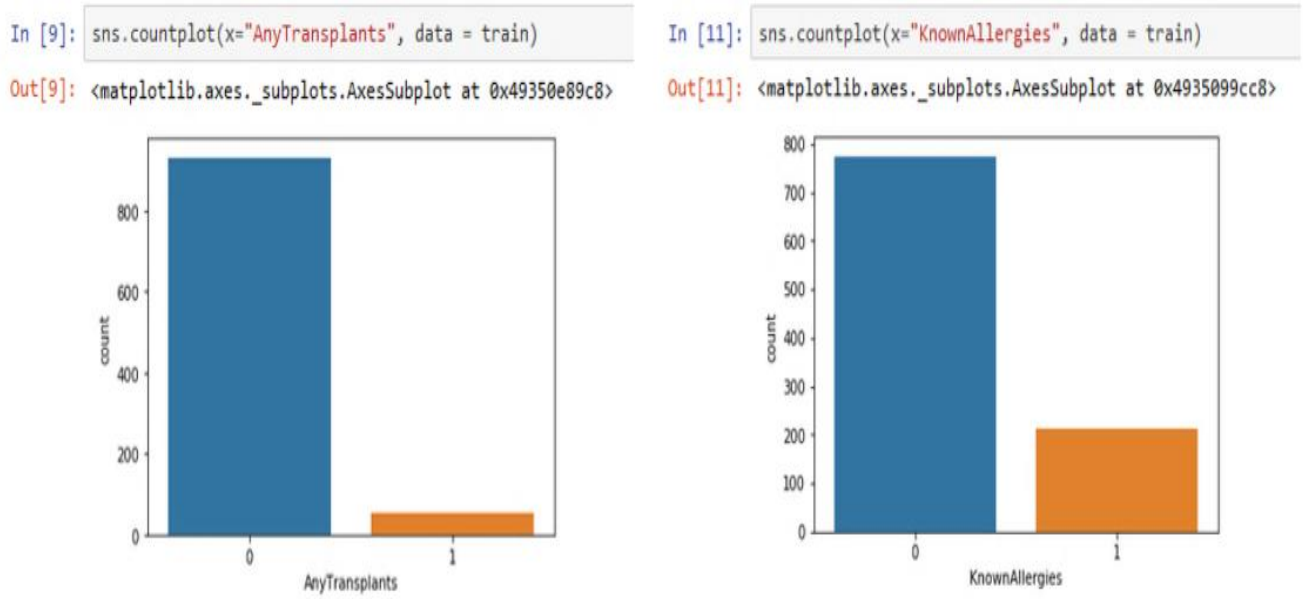


**Figure 2:** Plot on AnyTransplants and Known Allergies

Heatmap() function in seaborn library is used just to check the correlation between the features and the label.

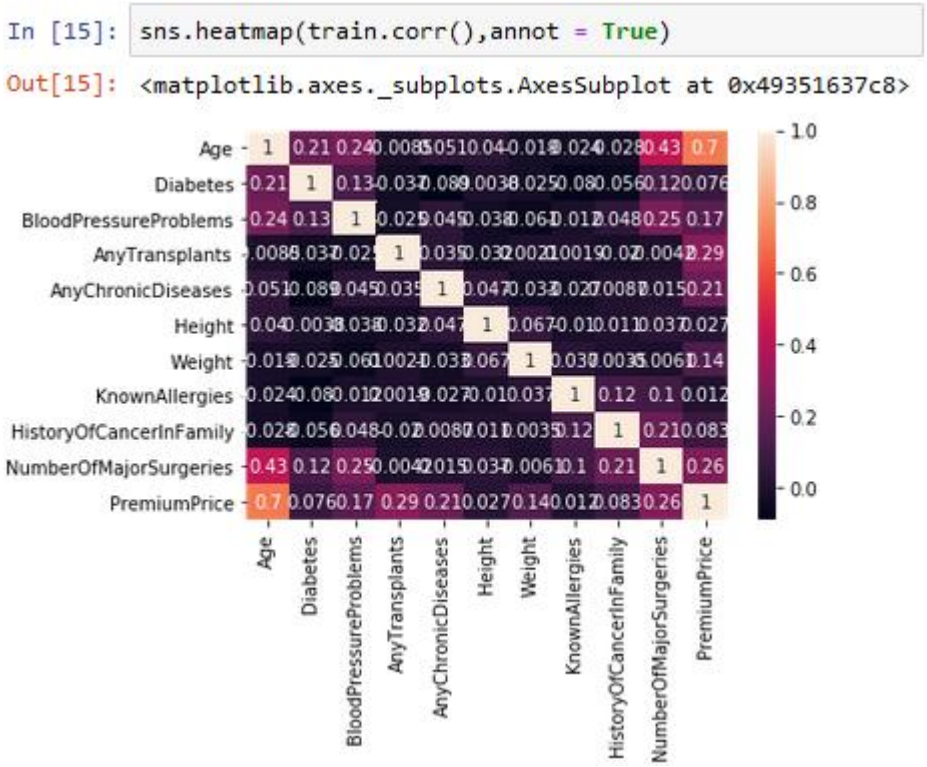


**Figure 3:** HeatMap on Various Attributes

**Splitting Into Training and Testing Dataset**

Here, scikit-learn library is most often used for the training and testing of the dataset where training dataset is 70% and testing dataset is of 30% which will be effective for further data preprocessing. By using sklearn model, we can import library train_test_split which is essential for the training and testing of the data. Training data is said to be an actual data, where as testing is an observed or selected data which helps for the model evaluation.

## 5. Results and Analysis

| Test | Test | Test | Test Steps | | | Test | Test |
|---|---|---|---|---|---|---|---|
| Case Index | Cases Operations | Cases Explanation | Step | Expectation | Final | Case Status | Priority |
| | Start the | Hosting the App ON and evaluate | If it | There is | The | Success | High |
| 01 | Application | Provoking the major requirement software must present. | isn't load | No place to Access | App is running successfully | | |
| 02 | Home | Check the Deployment Environment for properly loading the application deployment | If it isn't load | No access to an application | The App is running successfully | Success | High |
| 03 | Mode of User | Describe the Working of the app in freestyle mode | Whether not responding | No usage of freestyle methods | Displaying the freestyle application. | Success | Medium |
| 04 | Input of Data | Confirm whether the App can takes input and updates. | Failing to take an input or database storing. | No further proceedings | Updating of input towards specific routine. | Success | High |

**Test Results:** All the given test cases are successfully passed based on the specific criteria. There is no fault for the running of an application

## 6. Conclusion and Future Scope

Hence the study of deep learning through the Multi View framework helps to gain the simplification of the representation of the various patient medical expenses that can be predicted depending on the cause. Here, the perspective

support is on a FNN which is an attention based two directional RNN, and a layered attention network to dissimilate the mixed information in an electronically claim data depending on the various views. Practical results can be described our hail which defeats the various baselines on a day to day data of doctors who treats newborn and their expenses in each calendar year.

## References

[1] M. A. Morid, O. R. L. Sheng, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, "Healthcare cost prediction: Leveraging fine-grain temporal patterns," J. Biomed. Inform., vol. 91, 2019, Art. no. 103113.
[2] A. S. Ash et al., "Using diagnoses to describe populations and predict costs," Health Care Financing Rev., vol. 21, pp. 7–28, 2000.
[3] M. E. Cowen, D. J. Dusseau, B. G. Toth, C. Guisinger, M. W. Zodet, and Y. Shyr, "Casemix adjustment of managed care claims data using the clinical classification for health policy research method," Med. Care, vol. 36, pp. 1108–1113, 1998. 70 VOLUME 2, 2021
[4] A. K. Rosen, S. A. Loveland, J. J. Anderson, C. S. Hankin, J. N. Breckenridge, and D. R. Berlowitz, "Diagnostic cost groups (DCGs) and concurrent utilization among patients with substance abuse disorders," Health Serv. Res., vol. 37, pp. 1079–1103, 2002.
[5] D. Bertsimas et al., "Algorithmic prediction of health-care costs," Oper. Res., vol. 56, pp. 1382–1392, 2008.
[6] D. O. Clark, M. Von Korff, K. Saunders, W. M. Balugh, and G. E. Simon, "A chronic disease score with empirically derived weights," Med. Care, pp. 783–795, 1995, doi: 10.1097/00005650-199508000-00004
[7] M. Von Korff, E. H. Wagner, and K. Saunders, "A chronic disease score from automated pharmacy data," J. Clin. Epidemiol., vol. 45, pp. 197–203, 1992.
[8] P. A. Fishman, M. J. Goodman, M. C. Hornbrook, R. T. Meenan, D. J. Bachman, M. C. O'Keeffe Rosetti, "Risk adjustment using automated ambulatory pharmacy data: The RxRisk model," Med. Care, vol. 41, pp. 84–99, 2003.
[9] J. P. Weiner, B. H. Starfield, D. M. Steinwachs, and L. M. Mumford, "Development and application of a population-oriented measure of ambulatory care case-mix," Med. Care, pp. 452–472, 1991, doi: 10.1097/00005650-199105000-00006
[10] Y. Zhao et al., "Measuring population health risks using inpatient diagnoses and outpatient pharmacy data," Health Serv. Res., vol. 36, pp. 180–193, 2001.
[11] https://academic.oup.com/milmed/article/142/10/761/4898886?login=false
[12] https://dl.acm.org/doi/10.1145/2939672.2939785
[13] https://www.tandfonline.com/doi/abs/10.1080/10920277.2020.1754242?journalCode=uaaj20
[14] https://flask.palletsprojects.com/en/2.1.x/
[15] https://www.coursera.org/learn/neural-networks-deep-learning
[16] https://dl.acm.org/doi/10.1145/2750511.2750521
[17] https://ojs.aaai.org//index.php/AAAI/article/view/5400
[18] https://epubs.siam.org/doi/10.1137/1.9781611974973.23
[19] https://dl.acm.org/doi/10.1145/2939672.2939823
[20] https://www.tandfonline.com/doi/full/10.1080/10920277.2015.1110491
[21] https://github.com/keras-team/keras
[22] https://www.hindawi.com/journals/jhe/2022/7969220/
[23] https://www.researchgate.net/publication/303326261_Machine_learning
[24] https://www.cs.utexas.edu/~mooney/cs391L/paper-template.html
[25] https://omdena.com/blog/student-debt/
[26] https://www.dataquest.io/blog/data-science-portfolio-machine-learning/

## Author Profile

**Dr. Subhani Shaik** is Associate Professor in Sreenidhi Institute of Science and Technology. My Research Experience in Different fields like Data Mining; Machine learning, Artificial Intelligence and Data Science. More than forty publications in international and national journals and conferences and fifteen years of teaching experience. Act as reviewer for different Journals. Ph. D (CSE) received from Acharya Nagarjuna University. Master of Technology (CSE) from JNTUH, Hyderabad. Bachelor of Technology (CSE) from Andhra University, Visakhapatnam.

**Mr. P. Vinay Kumar,** final year students of information Technology, Sreenidhi Institute of Science and Technology (A), Hyderabad.

**Mr. Rohith Viswanathan,** final year students of information Technology, Sreenidhi Institute of Science and Technology (A), Hyderabad.

**Mr. M. Ganesh,** final year students of information Technology, Sreenidhi Institute of Science and Technology (A), Hyderabad.