Analysis of ICU Data of Pediatric Heart Patients

Sharmin Nahar Sharwardy[a],*, Hasan Sarwar[b], Mohammad Zahidur Rahman[a]
[a]Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh
[b]Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh
E-mail: sharmin34@yahoo.com

***Abstract*—One of the most fatal conditions for pediatric disease is congenital heart disease (CHD). Reviewing massive databases, comparing them, and mining them for information that can be used to identify, monitor, and treat illnesses like CHD is the key to treating cardiovascular disease. Cardiovascular disease can be predicted, prevented, managed, and treated with great effectiveness using big data analytics, which is well-known all over the world for its useful application in controlling, contrasting, and managing massive datasets. This study analyzes the post-operative ICU data. We analyzed the patient conditions of the ICU patients by using descriptive statistics. Here, we have selected the ICU parameters between different demographic groups by using chi-square test, t test and p value. Besides, we also used different machine learning methods to predict the patient's condition. The outcomes will serve as a reference for medical professionals employing big data technology to predict and manage CHD patients in ICU.**

***Index Terms*—CHD, Prediction, Machine Learning Methods, Big Data, Statistical Metrics.**

I. INTRODUCTION

In medical Field, different hospitals collect data, both structured and unstructured, about patients and their medication. The stored data can help the physician making better clinical decisions by reviewing the patient's previous health conditions and medication. "Big Data" or "unstructured data" refers to enormous data collections that cannot be handled, stored, or analyzed using traditional methods. It is still kept on record but not examined [1]. Due to the difficulty of searching and analyzing such data because of the lack of a clearly defined schema, it needs a particular technology and method to turn it into value [2]. As a result, big data analytics refers to the tools and methods used to analyze and collect data from massive amounts of data. Big Data analysis outcomes can be utilized to make future predictions. They also contribute to the development of historical patterns. Using data mining techniques, it is possible to analyze huge datasets derived from hundreds of patients, spot correlations and clusters, and create prediction models [3]. This paper analyzed and characterize the data set from post-operative ICU patient.

Physicians may manage data from a variety of sources. The patient's records are always being updated by healthcare practitioners with new information. This is done by medical professionals including doctors, nurses, and technicians in the form of reports or coded data referring to diagnoses or procedures. Additionally, a sizable amount of data produced by medical devices and systems is gathered, such as extracorporeal membrane oxygenation (ECMO) systems and dialysis monitor operating parameters as well as treatment protocols, drug administration databases, laboratory analyzer outcomes, vital signs, advanced monitoring data, and ventilator parameters. Each of this data was lost or, at most, stored as case histories on paper until a few years ago [4]. However, it is now possible to digitally store and interpret this data automatically, extract new knowledge from it, and use it to guide better patient care. Systems that can swiftly learn about the data produced by people in clinical care must be put in place in order to accomplish this goal. This will make it possible to make decisions based on data, improve forecasts for each patient's prognosis and treatment response, and gain a better knowledge of the complicated variables and how they interact to affect a patient's health.

Therefore, the purpose of this article is to identify the independent and dependent variables as well as predict the variables to figure out what the patient condition would be in the upcoming hour. Besides, to evaluate the statistical metrics of the machine learning algorithms to identify which algorithm is suitable for patients' Big data.

The remainder are arranged as follows in the article. Presenting notable connected studies is Section 2. Section 3 offers the dataset's specifics. The experiment's approach is described in Section 4. The result and analysis are presented in section 5. A summary of the study is included in the conclusion of Section 6.

II. RELATED WORKS

There is a great deal of research that demonstrates what potential data analysis can provide and what types of data can be evaluated. The disease-centered structure of healthcare management has given way to a patient-centered concept in recent years, even under value-based healthcare delivery models [2,5]. To provide good patient-centered care and uphold the requirements of this model, management and analysis of healthcare data are essential. Zoie et al used artificial intelligence for analytics big data [6]. They used different types of infection disease data. To anticipate and manage cardiac attacks, Chery et al. examined big data technology [7]. The optimum treatment strategy for heart attacks is suggested in this research using data mining, a crucial Big Data technology. For ICU patient big data analysis help to predict early mortality risk. Zeina Rayan et al described Machine learning technologies used to analysis ICU data [8]. Reiz et al. presented Big Data and Machine Learning and looked at potential clinical applications in the medical field [9]. Additionally, they explore potential optimization strategies for these new technologies as well as a new kind of hybrid healthcare-data science professional that serves as a link between physicians and data. Ana

Almeida et al [10] perspective on how to construct a system that can handle data in real-time, conduct analyses, and apply algorithms. Large amounts of data should be handled by this system, and through analysis and algorithms, it should be able to offer useful information. The present methodologies are also examined in this research along with how real-time operations and predictions can be made using them. The utilization of new technology in patient care and health management is now made possible by medical data analytics in the healthcare industry. The potential applications of medical data Analytics in healthcare were examined by Kornelia Batko et al. [11]. This essay's foundations include a critical examination of the relevant literature and the presentation of a few key findings from in-depth investigation into the application of big data analytics in healthcare settings. Electronic health records (EHRs) are becoming widely used, opening the door for big data research and bringing data science to the patient's bedside. The ICU makes a particularly strong argument for employing data science to enhance patient care within the healthcare system. L. Nelson et al [12] presented the definitions, types of algorithms, applications, challenges, and future of big data and data science in critical care. In order to examine comorbidity across various demographic groups and identify which comorbid health issues coexisted in patients with hospital-acquired pressure injury, Sookyung Hyun characterized comorbid conditions in intensive care unit (ICU) patients [13]. Giorgia Carra[14] provided a brief description of Big Data in the ICU. They also focused on different models face to reach the bedside and effectively improve ICU care. Many researchers have utilized deep learning techniques to investigate the potential of prediction systems. For example, [15,16], [17], and [18,19] have been developed recently to predict and risk analysis in ICU. Xiaobing Huang [20] described a deep learning model that can continuously track a patient's response to tranquilizer therapy, assess the treatment strategies of specialists to prevent dire circumstances like reverse medication associations, collaborate with a helpful mediator, and modify the strategies of experts as necessary. Muhammad Talha also focused on the integration between data set and deep learning in medical domain [21].

III. DATASET COLLECTION

The National Heart Foundation Hospital and Research Institute in Dhaka, Bangladesh provided the data set. The information from 410 ICU patients was gathered and examined. Patients aged 0 to 10 years were gathered for this investigation. The source dataset consists of over 25000 number of instances. Children under the age of 10 with congenital cardiac disease who were admitted to the ICU following post-operative surgery and had an ICU duration of stay of more than 24 hours met the inclusion criteria. The detail of the various features of the dataset is presented in Table 1.

Table I
Feature names, Ranges

Features name	Unit	Normal Ranges
Age		0-10years
Weight	Kg	
Heart Rate(HR)	beat/min	
Central venous pressure (CVP)	mmHg	2-6
HCO3	mmol/L	22-26
PH		7.5-7.45
Urine output	ml/hour	1-3

IV. METHODOLOGY

To conduct the study, we preprocessed the data before learning the models. The steps we took into consideration are described below-

A. Data Preprocessing

The raw data may consist of missing values, so we first removed the missing values from the raw dataset. Data Preprocessing is a crucial step in turning raw data into a format that can be understood. Preprocessing includes treatment of missing data and noisy data. Missing data arise when there are no data recorded for the variable in the observation [22]. It happens as a result of either technological or human error. The standard technique is used in this paper to handle missing values [23]. Statistical approaches created at an early stage of manipulating missing values are known as conventional procedures [22]. Numerous strategies, including deletion, ignoring, and Mean/mode Imputation [24], make use of the statistical premise in this procedure. In this case, Mean/mode Imputation was utilized to improve the outcome.

B. Feature Selection

The data sets presented in the previous section were analyzed using FS approaches to find promising feature sets. Performance metrics were recorded as ML models were trained using the generated feature subsets. The selection of features was done by statistical analysis. Python was used to execute the statistical analysis, and the chi-square univariate test was used to find the ICU data features that were statistically significantly different from one another. If the difference had a p-value of 0.05 or lower, it was taken as significant.

C. CHI-SQURE TEST

We can determine whether two random variables are independent using the chi-squared test. This indicates that there is no interaction between the probability distributions of the different variables. if there is an association between the variables. The standard description of the test is

• H0: Independent variables

• H1: Cut and are not independent variables.

D. T test

The purpose of the t-test is to determine whether there are any differences in the distribution of a numerical variable between several groups of a nominal variable. A null hypothesis (H0) and an alternative hypothesis (H1) are used to characterize the hypothesis test formally.

• H0: There are no variations in the variable distributions.

• H1: The variable distributions differ from one another.

E. Learning Models

In this phase, the preprocessed dataset is used to train seven different machine learning models described below. The goal of learning the models is to predict the target class of the patient from the dataset and find out what models fit the patients' Big data efficiently for making a decision.

Linear Regression: A dependent variable, a set of independent variables, and a linear relationship between the dependent and independent variables must be present for linear regression to be used statistically, which is primarily used for forecasting and predicting values based on historical data under certain key assumptions, that is:

$$y=a_1x_1+a_2x_2 ,\ldots+b+e \quad (1)$$

Where y is the response variable, e the error term, which has a normal distribution with mean zero and constant variance, and a and b the estimated regression coefficients [25].

Support Vector Machines: Although Support Vector Machines (SVM) is frequently thought of as a classification method, it can also be used to solve regression issues. Multiple continuous and categorical variables can be handled with ease. To divide various classes, SVM creates a hyperplane in multidimensional space. To reduce error, SVM builds an ideal hyperplane in a process that is iterative.

V. RESULTS AND DISCUSSION

We trained different models using Python and R library. All experiments were executed on Google Colab through the browser. The dependent and independent variables are taken from ICU doctors. Age,HR,PH,HCO3,BP,CVP,Urine output were selected for use in 42 attributes. The variable used in the data sheet was checked by an expert interview. The data were initially collected based on age. The patients are chosen whose age is 10 years or less. Here HR is heart rate, BP is the blood pressure, Urine is the hourly urine output, Central venous pressure (CVP) is an estimate of right atrial pressure, PH is the acidity level in the blood and HCO3 is the Bicarbonate rate. If any parameter exceeds the range, it may affect another parameter. Suppose, abnormal PH, HCO3 may cause abnormal HR. And abnormal HR may even cause sudden cardiac death. Table 3 shows the dependent and independent variables chi-test, p -value and t test among the variables. In the Table 2, P value of HR and HCO3 are less than 0.05, which reject Null Hypothesis (H0). That means, they are statistically significant.

Table 2
statistical analysis

Independent variables	Dependent variables	CHI-test	p-value	T -test
HR	HCO3	33.001	0.0002723	29.31
BP	PH	13.258	0.234	15.249
Na	HR	5.4885	0.4828	20.89
CVP	Urine	23.54	0.327	3.994

The linear regression model's residual plot, as depicted in Fig 1. The following provides an explanation of the linear regression model's output:

Standardized residual: The residual error has a constant variance and is located close to the zero line's mean.

QQ-plot: Any deviations are going to be skewed by the line, and all of the black dots in the QQ graph indicate the ordered distribution of the residuals. It is normally spread equally and follows N(0, 1).

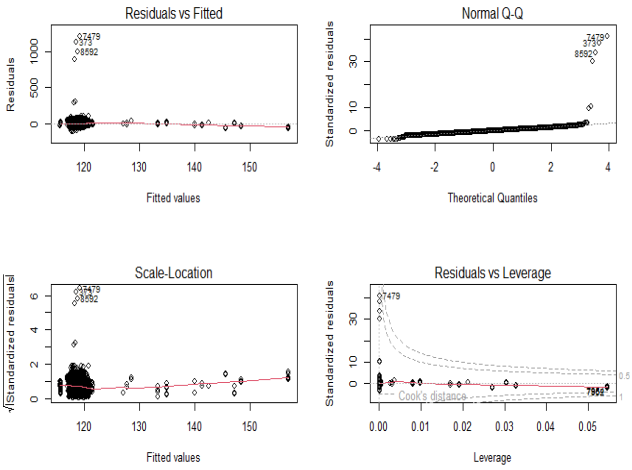

Fig.1.Residuals plot by linear regression.

The dataset is divided into training and testing portions that are 80%:20% for the linear regression model, respectively. To determine the precision of the prediction, we first contrast the predicted value with the time series' actual estimated value. Now that we have numbers and corresponding confidence intervals, we can also analyze and predict time series. To compare the actual value to the anticipated value in Fig. 2, we predict the test data. The training dataset is represented by the blue line in the image, while the predicted outcome is represented by the red line.

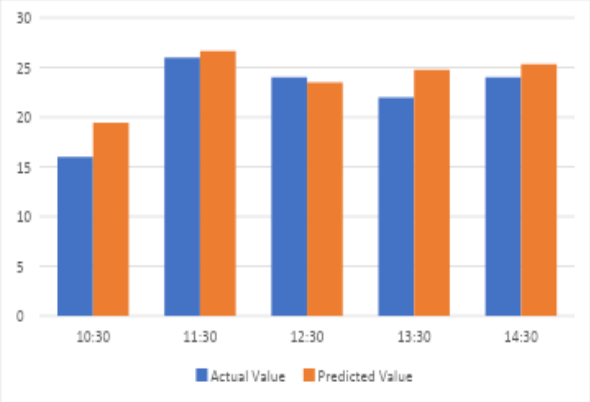

Fig. 2 Actual and expected value comparison chart

We also used test data to predict next hour patient condition using support vector machine algorithm. Let's utilize usual accuracy measures to evaluate the prediction outcomes; the outcomes are displayed in Table 3.

Table 3
Accuracy of Linear Regression Model and SVM Model.

Learning model	RMSE	MAE	MAPE
Linear Regression	0.1517	0.1204	0.009
Support vector machine	2.1015	0.1514	0.01012

For linear regression the Root Mean Squared Error is 0.1517, Mean Absolute Error is 0.1212 and Mean Absolute Percent Error is 0.009. For Support vector machine Root Mean Squared Error is 2.1015, Mean Absolute Error is 0.1514 and Mean Absolute Percent Error is 0.01012. For usage with the ICU pediatric congenital heart disease dataset in this study, the Linear Regression model is built to get superior efficiency than the Support vector machine model based on the results obtained from the table above.

CONCLUSION

Prediction of ICU data set is a big problem in Health sciences. Even if different techniques exist in this area, it is difficult to analyze noisy, irregular time series data set in ICU. The main goal of this study was to determine the ICU patient status of children with congenital cardiac disease. The study applies two different models to address this issue, namely linear regression and SVM algorithm on time series data, and analyzed their accuracy. Our future goal is to obtain data from more patients who are admitted to ICU with congenital heart disease and then evaluate it. Moreover, when the patient is in serious condition our solution does not work well, so we would like to work more to improve this portion as well.

ACKNOWLEDGMENT

The authors would like to thank the National Heart Foundation Hospital and Research Institute authority in Dhaka, Bangladesh for providing hospital patients with access to information. Authors also thank Bangladesh's Ministry of Information and Communication Technology for their financial support for the PhD fellowship.

REFERENCES

1. Raghupathi, W. and Raghupathi, Big data analytics in healthcare: promise and potential. *Health information science and systems*, *2*, pp.1-10.2018.
2. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. Journal of big Data. 2022 Jan 6;9(1)
3. Ekwonwune EN, Ubochi CI, Duroha AE. Data Mining as a Technique for Healthcare Approach. International Journal of Communications, Network and System Sciences. 2022 Nov 16;15(9):149-65.
4. Reiz AN, de la Hoz MA, García MS. Big data analysis and machine learning in intensive care units. Medicina Intensiva (English Edition). 2019 Oct 1;43(7):416-26.
5. Dash S, Shakyawar SK, Sharma M, Kaushik S. Big data in healthcare: management, analysis and future prospects. Journal of big data. 2019 Dec;6(1):1-25.
6. Wong ZS, Zhou J, Zhang Q. Artificial intelligence for infectious disease big data analytics. Infection, disease & health. 2019 Feb 1;24(1):44-8.
7. Alexander CA, Wang L. Big data analytics in heart attack prediction. J Nurs Care. 2017 Apr;6(393):2167-1168.
8. Rayan Z, Alfonse M, Salem AB. Intensive care unit (ICU) data analytics using machine learning techniques. Int J Inf Theor Appl. 2019;26(1):69-82.
9. Reiz AN, de la Hoz MA, García MS. Big data analysis and machine learning in intensive care units. Medicina Intensiva (English Edition). 2019 Oct 1;43(7):416-26.
10. Almeida A, Brás S, Sargento S, Pinto FC. Time series big data: a survey on data stream frameworks, analysis and algorithms. Journal of Big Data. 2023 May 28;10(1):83.
11. Batko K, Ślęzak A. The use of Big Data Analytics in healthcare. Journal of big Data. 2022 Jan 6;9(1):3.
12. Sanchez-Pinto LN, Luo Y, Churpek MM. Big data and data science in critical care. Chest. 2018 Nov 1;154(5):1239-48.
13. Hyun S, Newton C. Comorbidity analysis on ICU big data. The International journal of advanced culture technology. 2019;7(2):13-8.
14. Carra G, Salluh JI, da Silva Ramos FJ, Meyfroidt G. Data-driven ICU management: Using Big Data and algorithms to improve outcomes. Journal of critical care. 2020 Dec1;60:300-4.
15. Mikolov T, Chen K, Corrado G, Dean J. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781. 2013 Jan 16.
16. Wang M, Niu SZ, Gao ZG. A Novel Scene Text Recognition Method Based on Deep Learning. Computers. Materials & Continua.2019;60:781-94. doi: 10.32604/ cmc. 2019.05595
17. Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. EMNLP. 2014;14:1532-43
18. Aczon M, Ledbetter D, Ho L, Gunny A, Flynn A, Williams J, Wetzel R. Dynamic mortality risk predictions in pediatric critical care using recurrent neural networks. arXiv preprint arXiv:1701.06675. 2017 Jan 23.

19. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016 (pp. 770-778).

20. Huang X, Shan S, Khan YA, Salem S, Mohamed A, Attia EA. Risk assessment of ICU patients through deep learning technique: A big data approach. Journal of Global Health. 2022;12.

21. Khan M, Jan B, Farman H, Talha M, Ali S, Shah S, Khan FG, Iqbal J. Integration of big data and deep learning. Deep Learning: Convergence to Big Data Analytics. 2019:43-52.

22. Pratama I, Permanasari AE, Ardiyanto I, Indrayani R. A review of missing values handling methods on time-series data. In2016 international conference on information technology systems and innovation (ICITSI) 2016 Oct 24 (pp. 1-6). IEEE.

23. Rubin, Donald B., and J. A. Roderick. "Little, Statistical analysis with missing data." (1987).

24. Aydilek IB, Arslan A. A hybrid method for imputation of missing values using optimized fuzzy c-means with support vector regression and a genetic algorithm. Information Sciences. 2013 Jun 1;233:25-35.

Biography

Sharmin Nahar Sharwardy has completed her B.Sc and M.Sc in Computer Science and Engineering from National University, Dhaka, Bangladesh. She is doing her Ph.D.in the Department of Computer Science and Engineering at Jahangirnagar University in Savar, Dhaka. She has received ICT fellowship for Ph.D research in 2019 from Ministry of Information and Communication Technology , Bangladesh. Her research interests include Health informatics, Artificial Intelligence, Neural Networks and Biometric.

Hasan Sarwar has completed his B.Sc in Computer Science and Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh. He has completed his M.Phil. and Ph.D in Applied Physics from Electronics and Communication Engineering, University of Dhaka. He is now working as a Professor at the Department of Computer Science and Engineering, United International University, Dhaka, Bangladesh. His research interests include Software Engineering, Health Engineering, Education Data Mining, Bangla OCR, Nanomaterial Development.

Mohammad Zahidur Rahman has completed his B.Sc. and M.Sc Engineering in Electrical and Electronic Engineering from Bangladesh University of Engineering and Technology, Dhaka, Bangladesh in 1984 and 1990 respectively and has completed his Ph.D degree from University of Malaya, Kualalumpur, Malaysia. He is now working as a Professor at the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. His research field is E-Commerce, Computer Security, E-Governance, Communication, He has more than hundred research papers in national and international journals and conference proceedings.