IRJEMS International Research Journal of Economics and Management Studies
Published by Eternal Scientific Publications
ISSN: 2583 – 5238 / Volume 2 Issue 1 March 2023 / Pg. No: 333-344
Paper Id: IRJEMS-V2I1P143, Doi: 10.56472/25835238/IRJEMS-V2I1P143

Original Article

Machine Learning Models Powered by Big Data for Health Insurance Expense Forecasting

Jaya Vardhani Mamidala[1], Sunil Jacob Enokkaren[2], Avinash Attipalli[3], Varun Bitkuri[4], Raghuvaran Kendyala[5], Jagan Kurma[6]

[1]*University of Central Missouri, Department of Computer Science*
[2]*ADP, Solution Architect*
[3]*University of Bridgeport, Department of Computer Science*
[4]*Stratford University ,Software Engineer*
[5]*University of Illinois at Springfield, Department of Computer Science*
[6]*Christian Brothers University, Computer Information Systems*

Received Date: 23 March 2025 Revised Date: 02 April 2025 Accepted Date: 15 April 2025

Abstract: *Health insurance is the critical tool that can be adopted to reinforce the healthcare systems, especially among the low-income groups, and it does so by enhancing health outcomes, productivity and labor supply. Being aware of the cost of healthcare in terms of precise expense projections is important to policymakers and insurance agents. The research suggests a regime of machine learning to estimate health insurance expenses with the publicly available medical insurance cost prediction data hosted on Kaggle. The dataset consists of 2.7k records and their significant attributes such as age, BMI, and whether they smoke or not as well as their region. A thorough preprocessing procedure has been conducted; it includes data cleaning, removal of outliers, one-hot, and Z-score standardization. Gradient Boosting (GB) regression model was also applied in prediction of insurance expenses taking advantage of the ensemble learning behavior, which continuously minimizes errors in predictions. Predictive accurateness was high as the R 2 of the model was 92.0 and the Mean Squared Error (MSE) was given as 86.8. The outcomes confirm that Gradient Boosting is effective in fitting complex and non-linear relations that represent a viable source and mature scale solution to predicting personal healthcare expenditures.*

Keywords: *Health Insurance, Cost Prediction, Machine Learning, Healthcare Analytics, Policy Feedback, Insurance Expenses.*

1. INTRODUCTION

One of the most economic threats to an individual and family is health-related uncertainties that usually come at any time and without warning [1]. In contrast to discretional spending on house, cars, or other consumer durables which can be put off during bad times, medical needs require urgent care that consumes cash flow, and deeply interrupts savings and other life objectives of families like children's education, marriage, or retirement years, etc. [2]. Health insurance, in this case, is of great financial protection as it provides the needed healthcare services without causing economic crises and weakened financial situations. A well-performing health system is supposed to operate in four major dimensions to attain the following three core functions; health promotion, protection of financial risk and system responsiveness [3]. Healthcare Financing Strategy and similar policy efforts have been pushing for new types of social protection in underdeveloped nations, such as community-based health insurance and micro-health insurance. Nevertheless, these plans are not yet mature and are also prone to many implementation issues [4]. Also, patient satisfaction has become an important indicator of the quality of healthcare and is positively associated with the use of services and compliance with treatments.

Big Data and Machine Learning (ML) integration have presented transformative potential with regard to health insurance and healthcare financing during recent years [5]. Besides aiding the traditional mobile health (m-health) functionalities, the big data technologies are also relevant in advancing healthcare investments particularly in the face of aging populations and increased cost of healthcare [6]. Such technologies allow data-driven financial planning in order to align the rising healthcare expenses with tight budgets of the government, by implementing focused cost reduction plans by its stakeholders. Consumers' Health insurance options are also determined by various factors, of which the financial ones, in the form of the initial payment of premium, payments, and out-of-pocket spending limits, prevail the most [7]. But they have to keep in mind that the non-financial factors like selecting provider of healthcare, care continuity, etc. have much influence on decision-making as well. This knowledge of the said trade-offs allows one to predict more accurately the emerging patterns in insurance enrolments and consumer behaviour [8]. It is necessary to mention that significant part of the spending on healthcare is due to small fraction of the population which is commonly referred to as the high-need high-cost patients. Insurance models are being made so as to

cover this group against a disaster health expenditure, and are thus managing to combat illness-caused poverty and encouraging mutual assistance.

The processes of mobilizing and distributing resources to the public health services encompass how the resources are channeled, the configuration of delivery systems [9], compensation mechanisms to their providers and user incentives, are components of healthcare financing. It is a complicated procedure that includes revenue collection, pooling risk, purchasing strategy, and establishing the institutional framework enabling equity and protection of own funds [10]. Equity, efficiency, sustainability, and service quality are the key pillars of successful financing of healthcare. Machine Learning (ML) has also advanced the field of healthcare by improving its ability to forecast and make decisions [11]. One area that is gaining usage in ML techniques is drug creation, illness diagnosis, medical perception, remote surveillance and predictive analytics. ML models are the best at digging out patterns based on historical data, therefore, allowing evidence-based decision-making [12]. Such functions make ML an important feature to predict healthcare costs, create optimal insurance claims and push strategic healthcare policies.

A) Motivation and Contribution of the Study

The rising healthcare expenses are making proper forecasting of medical insurance an increasingly important issue for the insurance providers and policyholders alike. Conventional actuarial procedures are not able to incorporate the multidimensional nonlinear interdependence between specific qualities and healthcare spending. As the popularity of healthcare datasets used and machine learning algorithms evolve at an impressive rate, a potential possibility emerges to construct predictive models capable of providing highly accurate and individual forecasts. The purpose behind this study is anchored by the desire to realize such advantages, especially in the form of the Gradient Boosting algorithm to build a powerful model that can forecast health insurance costs. Insurance companies can use the suggested strategy to improve their risk stratification and pricing models by including the most important attributesage, body mass index (BMI), smoking status, and geographic regionand individuals can use the data for informed health coverage planning.

These are the main contributions of the current research:

- Relying on the available publicly available data from Kaggle that holds demographic and lifestyle characteristics to predict personal health insurance premiums.
- Conducted all steps of data preprocessing, such as Data cleaning and transforms, outlier detection, One-hot transforms and Z-score normalization.
- Created and trained a Gradient Boosting (GB) regression model using the same concept of ensemble-based learning as a powerful way of picking up nonlinear associations in the data.
- Evaluated the models by common measures of regression e.g. coefficient of determination (R2) and MSE and confirmed applicability of the model in providing an honest assessment and subsequent predictive capabilities.

B) Justification and Novelty

The rationale of this study lies on the increased need of precise data-oriented forecasting models in health insurance industry where there are significant challenges in the application of conventional statistical methods due to their limited capability to capture complexity in the expenditure data in terms of nonlinearity. The proposed application of a Gradient Boosting (GB) regression model is a new contribution that will merge the most advanced ensemble learning methods with a well-organized preprocessing pipeline that is specific to the healthcare-related cost prediction task. This method works well with feature interaction, data variability and noise as compared with traditional models, making it a more predictable approach. The first of its kind is its entire reproducible end-to-end framework, which consists of data preprocessing in the real world, illuminating visualization, and solid model evaluation, which can be generalized and applied in healthcare analytics and insurance risk modeling in the future.

C) Structure of the Paper

The paper is organised as follows: Provide a literature review on health insurance cost forecasting in Section II; detail the technique in Section III; provide the results and model comparisons in Section IV; draw conclusions and propose future research topics in Section V.

II. LITERATURE REVIEW

In this section, the study reviews existing literature on forecasting health insurance using machine learning techniques. Key themes from the reviewed works include:

Akbar et al. (2020) Fraud involving healthcare insurance accounted for a 10% increase in yearly health expenditure, which amounted to $100 billion annually. One uses the current scientific knowledge to find and prevent fraud. This research aims to conduct an analysis of statistical modelling approaches for evaluating false health benefits utilising cutting-edge techniques. Once data collection and exploratory analysis are finished, it can find the best models using RF regression and the

classification of trees technique with XGB. Comparatively, XGB Tree obtained 87% accuracy with fraudulent suppliers and 86% overall by random sub-sampling, while RF only managed 81% accuracy with class 1 recall. Applying the XGB approach to clean, heavily adjusted data yields better results, as seen by the results [13].

Mahardini and Dachyar (2020) intended to enhance the hospital's claim fulfilment process through the use of BPR and MIS to shorten the time it takes to submit claim files and reduce the number of times claims are returned. Two parts make up business process reengineering: modelling and simulation of existing processes and modelling and simulation of future processes. Out of this research, four potential ways to streamline the public insurance claim procedure were developed, each with its own estimated processing time. An ideal situation would be one that is created using a data flow diagram and an ERD. The best combination of the highlighted scenarios is available in public insurance organizations' EMR, HIS, and enhanced claim fulfillment systems. With a 78.73% improvement in efficiency, the average cycle time of business processes was found to decrease from 28 days to 14 days in this study [14].

Laagu and Setyo Arifin (2020) express their thoughts on the growing cost of social security, also known as BPJS Kesehatan, using online social media. Being a social media platform that offers in-depth discussion and opinions, Twitter is its primary focus. They employed Drone Emprit, a big data system that records and analyses emotions and engagement on Twitter and other social media platforms, for their study of social media data. The data set being examined contains sixty days' worth of talks, starting from September 22, 2019, and ending on November 22, 2019 Out of 360,820 contacts that took place during that time, 91% were negative, 5% were positive, and 4% were neutral. Data from BPJS Kesehatan shows that the majority of Indonesians are opposed to the planned increase in national health insurance rates [15].

Zhu, Wu and Wang (2019) A neurological disease called epilepsy causes people to have seizures over and over again. About 0.5 to 1% of the world's population has this disorder. Almost a thousand persons with epilepsy die each year from Sudden Unexpected Death in Epilepsy (SUDEP), an illness that sadly has little public understanding. The purpose of this research is to examine the relationship between the mortality rate of epileptics and variables such as demographics, financial information, and identification codes. Using commercial insurance claims data from the United States and its associated diagnostic codes and non-diagnosis variables, they developed a mortality prediction model. Because they used different feature vectors, the classification accuracy they presented was 91.0 and 85.0, respectively. They built on their analysis of the aforementioned elements in the prediction model to include causal inference between altered diagnostic codes and chosen non-diagnosis factors [16].

Rao and Clarke (2018) used this model to extrapolate potential patient expenditures for various medical procedures using publicly available healthcare records. Investigators in New York State honed focused on over 2 million de-identified patient records stored in the SPARCS database. Models such as regression trees, multivariate linear regression, and DNNs were also considered. For determining the optimal parameter values, they employed a grid-search technique. The ideal setup was found to be an 8-layer DNN with dimensions 5x5x10x25x25x10x5x5 and an Adam optimizer with a learning rate of 0.01 using the commonly used R2 measure in grid-search cross validation. With an R2 value of 71.0, they outperformed previous reports of comparable issues in the literature [17].

Peters and Maxemchuk (2017) The goal was achieved by establishing a system of six autonomous processing units that could collaborate on a single task while maintaining the confidentiality of the patient's personal information, medical records, and demographic data. They compared the two programs' performance using Shannon entropy, which takes into account both the total amount of data stored in a system and the amount of data exposed by compromised components. One of the metrics evaluated was the amount of data revealed by compromised components. When it comes to clearinghouse assaults, this distributed strategy greatly decreases the quantity of entropy by approximately 79%. The MyPHRMachines architecture was also compared to their method. Based on their findings, it seems that either the distributed or centralized systems have a faulty component [18].

Peng and You (2016) provided a fresh approach to identifying health insurance fraud using neural networks. An MPL neural network with an analytical contribution hierarchy process inside it was used to build a pharmacopoeia spectrum tree, which grouped medical items using a neural network's analytical contribution hierarchy process. Then, using the method of hierarchy contribution rate and multidimensional space distance, the process determined the medical items' contribution rates of fraud. Lastly, this process was used to identify the medical items most likely to be fraudulent. The experiment findings showed that the enhanced neural network outperformed prior unsupervised data mining methods with an accuracy of 86% in detecting healthcare fraud [19].

The comparative analysis of background study based on their Methodology, Problem Addressed, Performance and Future Work is provided in Table I

Table 1: Review of Literature on Forecasting Health Insurance Expenses

Author	Methodology	Dataset	Problem Addressed	Performance	Future Work
Akbar et al. (2020)	Random Forest, Classification Trees, Extreme Gradient Boost (XGB), data tuning	Health insurance claims data	Detect fraudulent healthcare benefit claims	XGB: 86% overall accuracy, 87% accuracy with illegitimate providers; RF: 81% class 1 recall	Improve data tuning and explore additional ML techniques
Mahardini & Dachyar (2020)	Tools for Business Process Reengineering (BPR), MIS, ERD, and DFD	Hospital internal process data	Improve claim fulfillment time and reduce file return rate	Efficiency improved by 78.73%; claim process time reduced from 28 to 14 days	Further digital integration (EMR, HIS); broader application in other institutions
Laagu & Arifin (2020)	Sentiment analysis using Drone Emprit (big data system), social media analytics	Twitter data (22 Sept – 22 Nov 2019, 360,820 tweets)	Analyze public sentiment toward rising BPJS (Indonesia) national health insurance fees	91% negative sentiment, 5% positive, 4% neutral	Deeper topic modeling, expand to other social media platforms
Zhu, Wu & Wang (2019)	Classification models using diagnosis & non-diagnosis features; causal inference	U.S. commercial insurance claims	Predict epilepsy patient mortality (SUDEP)	Accuracy: 91.0% (full features), 85.0% (selected vectors)	Extend causal inference for diagnosis codes and non-diagnosis features
Rao & Clarke (2018)	Using a grid search, an 8-layer deep neural network (DNN), regression trees, and linear regression	SPARCS (NY state patient data; 2M+ records)	Predict costs of medical procedures from open healthcare data	$R^2 = 71.0$ with 8-layer DNN using Adam optimizer	Expand to national datasets; domain-specific cost prediction
Peters & Maxemchuk (2017)	Distributed processing of CMS-1500 form using autonomous processing units; entropy analysis	Conceptual data on CMS-1500 form, MyPHRMachines, simulated systems	Protect insurance claim data privacy using distributed architecture	79% reduction in entropy loss in distributed vs centralized system	Further development of distributed architectures for secure claim handling
Peng & You (2016)	Clustering using a spectrum tree; an Analytic Contribution Hierarchy Process-modified MPL neural network	Medical insurance claim data	Detect fraud in medical insurance by identifying suspect medical items	Accuracy: 86%, outperforming other unsupervised data mining methods	Refine fraud detection factors and apply method to broader fraud categories

III. METHODOLOGY

The proposed methodology for forecasting health insurance expenses involves a structured machine learning pipeline. The first step is to get the Medical Insurance Cost Prediction dataset from Kaggle. This dataset contains important personal and lifestyle variables including age, BMI, level of smoking, and location. The dataset is next under systematic preprocessing such as data cleaning to improve the absence of values or inconsistencies, outliers' identification and elimination to reduce distortion and the one-hot encoding process to convert categorical features to numerical references and the Procedure of normalization of Z-scores to normalize dimensions of features. Partitioning the dataset into training 80% and testing 20% groups helps with model validation after preprocessing. Gradient The complicated non-linear connections between insurance features and premiums can be learnt using a boosting regression model. Statisticians use the Coefficient of Determination (R^2) and the MSE to determine the model's efficacy and the robustness of the suggested approach. The main flow of the proposed system is shown in Figure 1.

A) Data Collection

This Endeavor makes use of the Kaggle Medical Insurance Cost Prediction dataset. Data from the medical insurance dataset may include demographic information that influences healthcare spending, such as age, sex, BMI, smoking status, family size, and location. Machine learning algorithms are trained on this dataset to forecast the medical expenses of incoming

policyholders. There are 2,700 rows and 7 columns in the dataset, and the fields included are age, body mass index, and charges. This section displays the data visualizations:

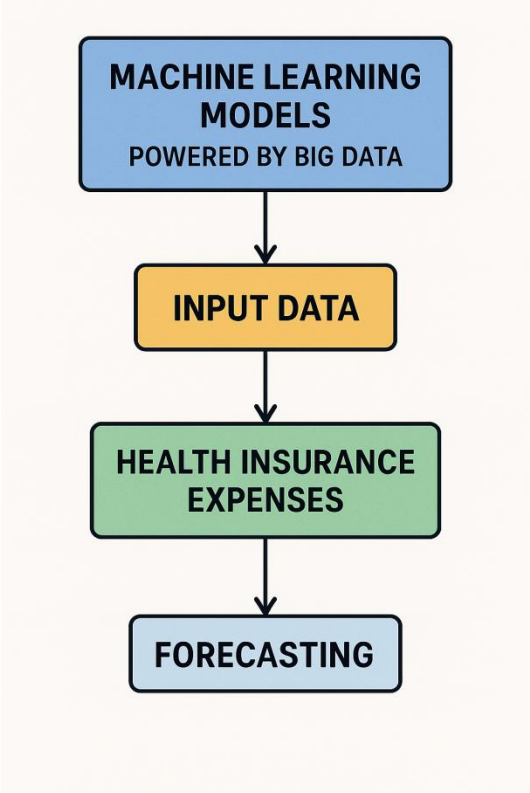

Figure 1: Flowchart Diagram of the Health Insurance Expenses

The overall steps of the flowchart for Health Insurance Expenses are explained below:

The Pearson correlation coefficients between the variables "age," "bmi," and "charges" are shown in Figure 2's heatmap. There is perfect self-correlation when the diagonal values are 1.00. The 'age' and 'charges' variables are positively correlated with one another at a moderate level (0.53), with a weak positive correlation at 0.27, and with a weak positive correlation at 0.23 between the 'age' and 'bmi' variables. Visually, the colour scale emphasizes these associations; greater correlations are indicated by red.

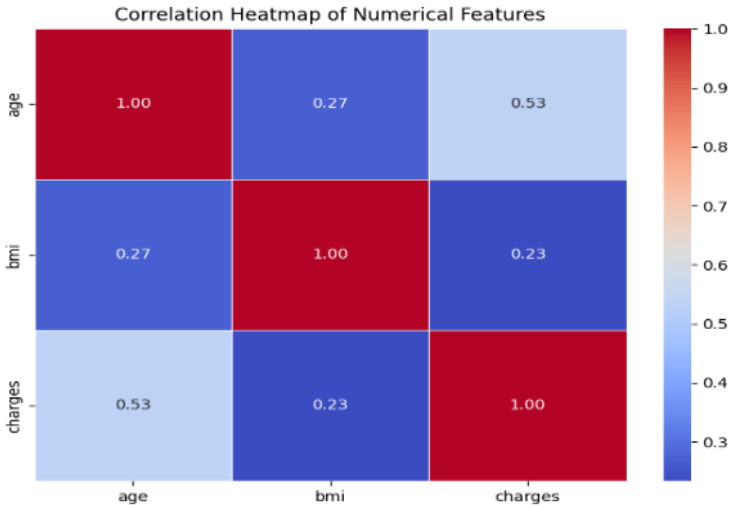

Figure 2: FiureCorrelation Heatmap of Numerical Features

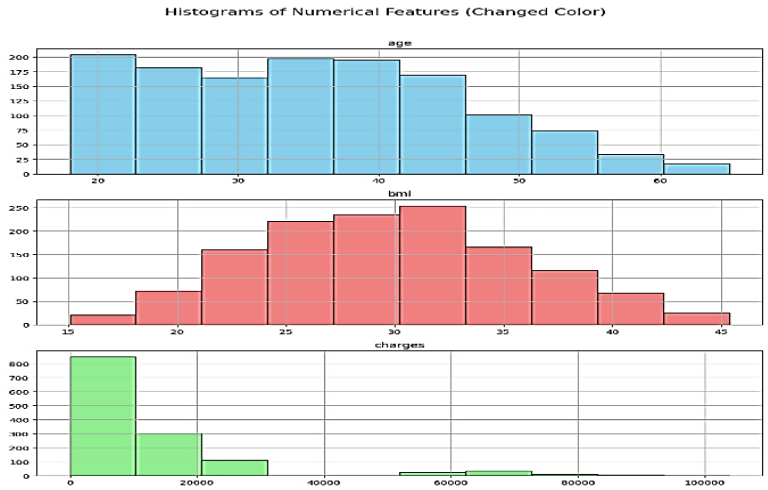

Figure 3: Distribution of Categorical Features

Figure 3 shows histograms for 'age', 'bmi', and 'charges'. The 'age' distribution is fairly uniform, with peaks in the early 20s and mid-30s, spanning 18 to 65 years. 'BMI' has a slightly right-skewed, bell-shaped distribution, mostly between 25 and 35, indicating many individuals are overweight or obese. 'Charges' is highly right-skewed, with most values below 10,000 and a few high-cost outliers creating a long tail.

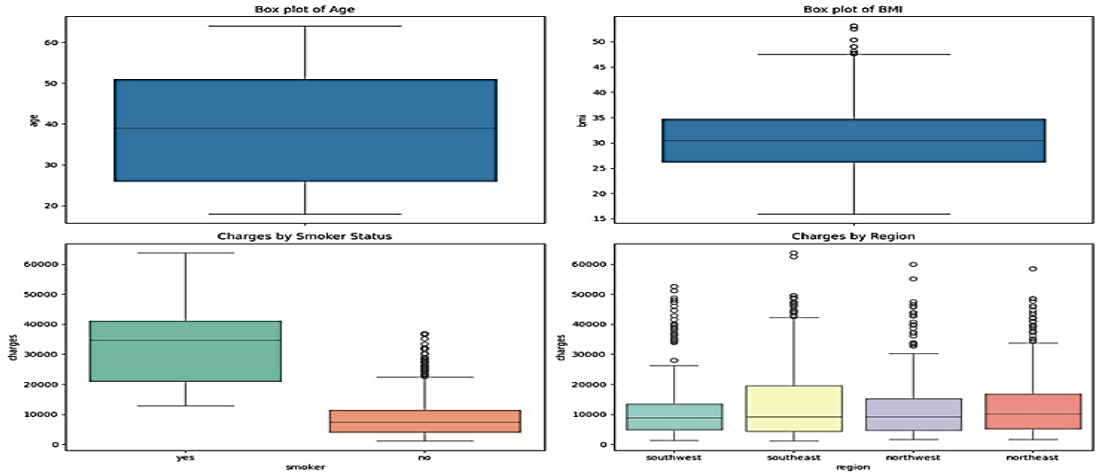

Figure 4: Box-Whisker Plot of Features

Figure 4 presents four box plots summarizing dataset attributes. The "Age" plot shows most individuals are aged 28–50, with a median around 40. The "BMI" plot reveals values mostly between 26–35, with outliers nearing 50. "Charges by Smoker Status" highlights significantly higher and more variable charges for smokers (median 35,000) versus non-smokers (10,000). "Charges by Region" shows similar median charges across regions, though the southeast exhibits a slightly higher median and a broader range with more outliers.

B) Data Preprocessing

The accuracy and dependability of machine learning models are directly affected by data preprocessing, which is an essential step in ensuring high-quality data prior to model training. The preprocessing workflow in this study involves several key steps likely data cleaning and transformation, one-hot encoding, managing outliers and data standardization using Z-score.

These stages are briefly discussed below:

a. Data Cleaning and Transformation

The initial statistics show that different age groups of workers, couples, and infants have different premium amounts. This data is converted into columns that show the duration, number of members in the family, and premium amount [20]. Age, family size, premium, and couple were the only remaining columns after preprocessing with column headers. The

original data included 36 columns. In these columns, you may find premium adults ranging in age from 21 to 27, couple+2, children in their 30s, and so on.

C) Managing Outliers

Clustering outliers in the residual histogram preference space is possible through outlier detection [21]. Most outliers in a dataset may be found using the outlier detection procedure; when those outliers are removed, the remaining dataset is subjected to inlier segmentation.

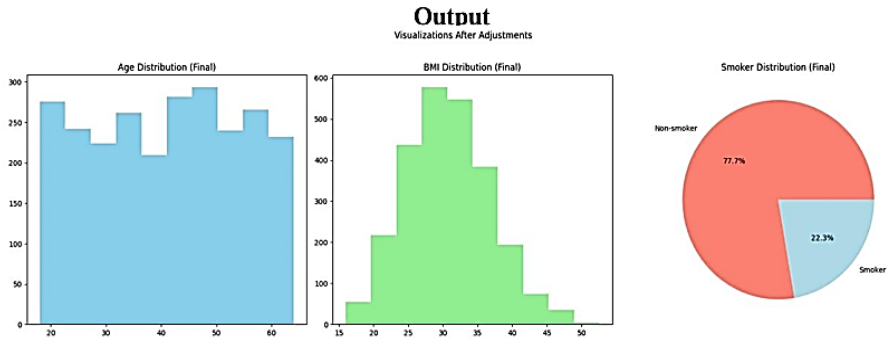

Figure 5: Visualization After Removing Outliers

Figure 5 displays three final distribution plots. The 'Age' histogram (light blue) shows a fairly uniform spread from ages 20 to 60, indicating diverse age representation. The 'BMI' histogram (light green) is unimodal, centered around 25–35, reflecting a majority in the healthy to overweight range. The 'Smoker' pie chart reveals a clear imbalance, with non-smokers (red) comprising 77.7% and smokers (light blue) 22.3% of the dataset.

D) One-Hot Encoding

The one-hot encoding format is used to express categorical information. With the One-Hot encoding, a word or character is represented by a vector with a single element that is one and zero for all the others. In a text-to-vector mapping, the size of the vocabulary determines the length of the vector x, and the element's location in the vocabulary determines which element is set to one.

E) Data Standardization using Z-score

The pre-processing step of standardization, or z-score normalization, is essential. This research makes use of a dataset that includes a wide variety of potentially linked continuous features, each with its own unique set of units and value ranges. Through the process of standardization, the continuous characteristics are brought to a uniform scale, where the mean is zero and the variation is one. One way to convert normal variates to a standard score is to use the following formula. It is the z-score formula that may be found in Equation (1):

$$Z_1 = \frac{X - \bar{X}}{\sigma} \qquad (1)$$

A sample's standard deviation (σ) and its mean ($\bar{X}$) are defined in relation to the original data value (X).

F) Data Splitting

An 80:20 split is used to partition the dataset, with 80% going into training the model and 20% into testing to see how well it performs.

G) Classification with a Gradient Boosting Model

A popular alternative to traditional Boosting techniques is Gradient Boosting (GB). This method of learning defines a strong classifier by gradually fitting new models [22]. In order to construct the first learner and fit each succeeding tree to the pseudo-residues of the prior tree's prediction, the new model iteratively minimizes the loss function. You may see the resulting Equation in (2) down below:

$$F(x) = G_0 + \beta_1 T_1 + \beta_2 T_2 + \cdots + \beta_n T_n \qquad (2)$$

G_0 is equal to the first value in the series, T_1. The algorithm finds the β_n coefficients for the tree nodes, and T_(n) are the trees fitted to the pseudo-residuals. Setting the number of estimators and boosting phases to run is an option you can configure. Values in the tens to hundreds of orders of magnitude are the most common for estimators. Set the number of estimators and boosting phases to run as you like. [23]. Typically, estimators are used with an order of magnitude ranging from

tens to hundreds. Both speed and accuracy are enabled by the GB classifier. The selection of weak learners or an excessive number of them can lead to overfitting, the most common issue with these estimators.

H) Performance Matrix

Classifiers constructed using various ML techniques and resampling strategies were evaluated using a number of metrics [25]. Machine learning models for classification and regression are evaluated using metrics like MSE and R2:

i. Coefficient of Determination (R2)

The explanatory goal of using an R2 regression metric is to show how well the expected and actual output values match up with one another. The MSE and the square of the variation in the Y-values (the denominator) form the basis of the formula for its calculation. This information is provided in Equation (3):

$$R^2 = 1 - \frac{\frac{1}{n}\sum_{i=1}^{n}(Y-\hat{Y})^2}{\left(\frac{1}{n}\sum_{i=1}^{n}(Y-\bar{Y}^2)\right)} \quad (3)$$

Where, Y is actual value, $\hat{Y}$ is predicted value and $\bar{Y}$is mean value.

ii. Mean Squared Error (MSE)

Instead of using the absolute value as in formula (4), minimal squared error (MSE) squares the difference between actual and expected output before adding them all:

$$MSE = \frac{1}{n}\sum(Y-\bar{Y})^2 \quad (4)$$

Where, Y= actual value, and $\bar{Y}$= predicted value.

The R^2 metrics has been used to make comparative analysis and rate the performance of the model on the Forecasting Health Insurance Expenses.

IV. RESULTS AND DISCUSSION

A local system with a 3600 6-Core Processor (3.60 GHz) and 16.0 GB RAM was more than able to manage the task, according to the trial setup. As shown in Table II, the Gradient Boosting (GB) model was able to successfully predict future health insurance premiums. As shown by the high coefficient of determination (R^2) of 92%, the model effectively accounts for a substantial amount of the variation in the dependent variable. An MSE of 86.8 was likewise attained by the model, which is indicative of a reasonably modest average disparity between the predicted and actual values. Results like these demonstrate that Gradient Boosting is an effective method for collecting complex data relationships and making reliable predictions about health insurance prices.

Table 2: Results of Gradient Boosting for Forecasting Health Insurance Expenses

Performance Metric	GB model
R2	92.0
MSE	86.8

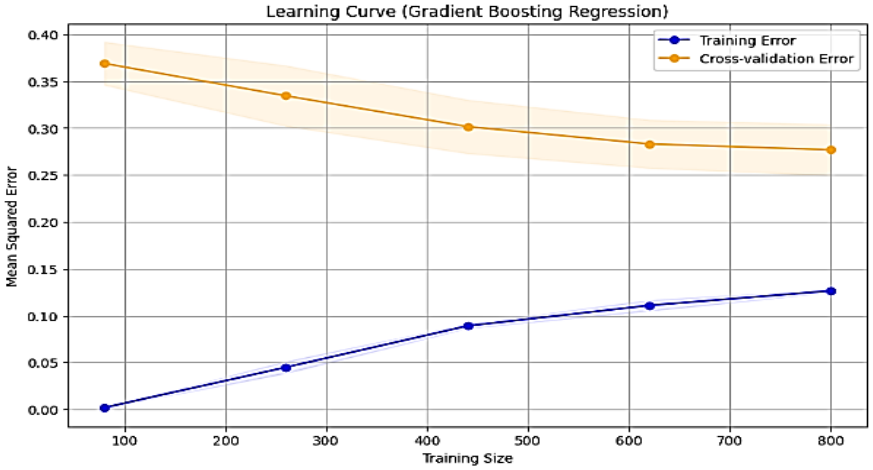

Figure 6: Learning Curve of Gradient Boosting Model

Figure 6 shows a Gradient Boosting Regression model's performance with increasing training data. The Training Error (blue line) rises, while the Cross-validation Error (orange line) generally decreases, both measured in Mean Squared Error. This indicates improved generalization with more data, though with diminishing returns after 600 samples. The gap between the two errors, especially at smaller training sizes, highlights initial high variance (overfitting) that lessens with more data, as further evidenced by the standard deviation shaded areas.

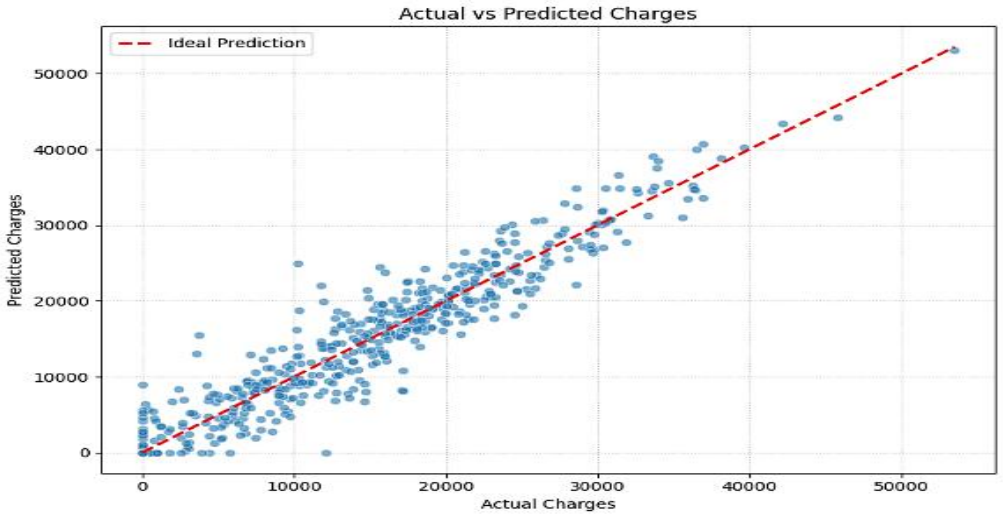

Figure 7: Actual vs. Predicted Charges

Figure 7 visualizes the regression model's performance by plotting actual charges against predicted values. Blue dots represent individual predictions, while the red dashed line ($y = x$) Indicates ideal predictions. Most points cluster near this line, showing strong predictive accuracy. A positive correlation is evident, though some dispersion at higher charge values suggests increased prediction error in that range.

A) Comparative Analysis

This section compares the proposed GB model to three popular machine learning algorithms: SVM and LR. The performance of the models based on their R-squared was summarized in Table III. GB model ranked no. 1 out of all models with R^2 equal to 0.920 which sufficiently shows its robust predictive power on the basis of health insurance forecasting. Comparatively, SVM model gave an R^2 90.0 . LR proved to have poor performance with 40.2 R^2 rating as nails do not satisfy this task. These findings indicate the success of machine learning models, which is due to GB in predicting health insurance expenses correctly.

Table 3: Comparative Analysis of ML Models for Forecasting Health Insurance Expenses

Models	R2
SVM [24]	90
LR [25]	40.2
GB	92.0

A high R^2 score of 92.0 indicates that the suggested model based on Gradient Boosting (GB) is a good predictor of health insurance premiums. This model stands out because to its exceptional performance in capturing intricate, nonlinear patterns in the data. The major strength of the GB model is that it enjoys the ensemble learning design that integrates the power of weak learners into strong predictor that makes them very strong in terms of accuracy and generalizability. Also, GB is insensitive to overfitting and effectively processes feature interactions, rendering it especially suitable in forecasting healthcare expenditures in the real world where variations and noisiness in the data are typical.

V. CONCLUSION AND FUTURE SCOPE

Healthcare funding is a basic issue in different regions, where there is considerable difference in evident coverage, hospital quality, and investment amount. Sustainable health insurance cost forecasting plays an essential role in policy creation, risk control, and sustainability in the scope of healthcare provision. The complexity of such work is in the variety of personal, lifestyle, and demographic factors, that affect medical expenditures. Machine learning is one of the strong methods to solve this challenge to find complex patterns in the data. The paper indicated that Gradient Boosting (GB) regression model has efficacy in predicting health insurance costs. The model represented the nonlinear interaction between leading variables and provided a high

level of predictive performance, which is evidenced by a substantial value of R 2 and a very low value of Mean Squared Error (MSE). Embarking on these findings indicates the promising uses of GB as a source of assuring the insurance cost reimbursement to assist evidence-based decision-making by the health insurance providers and policymakers.

To proceed in future, the model may be improved by adding other set of variables like medical history, claim rate or available healthcare service in order to better predict the model. More experiments using new complex ensemble algorithms such as XGBoost or CatBoost, and deep learning algorithms such as neural networks might provide better performance. The implementation of Explainable AI (XAI) approaches could introduce clarity to the model outcomes, making them more interpretable as well as boosting credibility among the stakeholders. Lastly, to evaluate the future research of the model, applying it to a larger amount and a more diverse set of data would be useful in determining the scalability and viability of the model in other populations and insurance markets. “

VI. REFERENCES

[1] R. Singh and A. Singh, “A Study of Health Insurance in India,” *Int. J. Manag. IT Eng.*, vol. 10, no. 4, pp. 558–2249, 2020.

[2] S. H. Zolfani, R. Dehnavieh, A. Poursheikhali, O. Prentkovskis, and P. Khazaelpour, “Foresight Based on MADM-Based Scenarios’ Approach: A Case about Comprehensive Sustainable Health Financing Models,” *Symmetry (Basel).*, vol. 12, no. 1, p. 61, Dec. 2019, doi: 10.3390/sym12010061.

[3] A. R. Sarker *et al.*, “Determinants of enrollment of informal sector workers in cooperative-based health scheme in Bangladesh,” *PLoS One*, 2017, doi: 10.1371/journal.pone.0181706.

[4] M. A. Hanifi *et al.*, “Profile: The Chakaria Health and Demographic Surveillance System,” *Int. J. Epidemiol.*, vol. 41, no. 3, pp. 667–675, Jun. 2012, doi: 10.1093/ije/dys089.

[5] W. Raghupathi and V. Raghupathi, “Big data analytics in healthcare: promise and potential,” *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, Dec. 2014, doi: 10.1186/2047-2501-2-3.

[6] J. Archenaa and E. A. M. Anita, “A Survey of Big Data Analytics in Healthcare and Government,” *Procedia Comput. Sci.*, vol. 50, pp. 408–413, 2015, doi: 10.1016/j.procs.2015.04.021.

[7] T. G. McGuire, “Demand for Health Insurance,” *Handb. Heal. Econ.*, vol. 2, pp. 317–396, 2011, doi: 10.1016/B978-0-444-53592-4.00005-0.

[8] A. J. Trujillo, F. Ruiz, J. F. P. Bridges, J. L. Amaya, C. Buttorff, and A. M. Quiroga, “Understanding Consumer Preferences in the Context of Managed Competition,” *Appl. Health Econ. Health Policy*, vol. 10, no. 2, pp. 99–111, Mar. 2012, doi: 10.2165/11594820-000000000-00000.

[9] G. F. Anderson *et al.*, “Attributes common to programs that successfully treat high-need, high-cost individuals,” *Am. J. Manag. Care*, 2015.

[10] E. Owusu-sekyere and D. A. Bagah, “Towards a Sustainable Health Care Financing in Ghana : Is the National Health Insurance the Solution ?,” vol. 4, no. 5, pp. 185–194, 2014, doi: 10.5923/j.phr.20140405.06.

[11] W. H. O. (World H. Organization), “Fifty-Eighth World Health Assembly,” *Wha58/2005/Rec/1*, no. May, pp. 1–159, 2005.

[12] S. D, S. V, and J. D, “Application of Machine Learning Techniques in Healthcare,” 2020, pp. 289–304. doi: 10.4018/978-1-5225-9902-9.ch015.

[13] N. A. Akbar, A. Sunyoto, M. Rudyanto Arief, and W. Caesarendra, “Improvement of decision tree classifier accuracy for healthcare insurance fraud prediction by using Extreme Gradient Boosting algorithm,” in *2020 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*, IEEE, Nov. 2020, pp. 110–114. doi: 10.1109/ICIMCIS51567.2020.9354286.

[14] S. R. Mahardini and M. Dachyar, “The critical improvement of hospital claim fulfillment towards public insurance, using BPR and MIS approach,” in *Proceedings of ICAE 2020 - 3rd International Conference on Applied Engineering*, 2020. doi: 10.1109/ICAE50557.2020.9350551.

[15] M. A. Laagu and A. S. Arifin, “Analysis the Issue of Increasing National Health Insurance (BPJS Kesehatan) Rates through Community Perspectives on Social Media: A Case Study of Drone Emprit,” in *2020 International Conference on Smart Technology and Applications (ICoSTA)*, IEEE, Feb. 2020, pp. 1–7. doi: 10.1109/ICoSTA48221.2020.1570615599.

[16] Y. Zhu, H. Wu, and M. D. Wang, “Feature Exploration and Causal Inference on Mortality of Epilepsy Patients Using Insurance Claims Data,” in *2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 2019, pp. 1–4. doi: 10.1109/BHI.2019.8834638.

[17] A. R. Rao and D. Clarke, “A comparison of models to predict medical procedure costs from open public healthcare data,” in *Proceedings of the International Joint Conference on Neural Networks*, 2018. doi: 10.1109/IJCNN.2018.8489257.

[18] E. Peters and N. Maxemchuk, “A Privacy-Preserving Distributed Medical Insurance Claim Clearinghouse & EHR Application,” in *Proceedings - 2017 IEEE 2nd International Conference on Connected Health: Applications, Systems and Engineering Technologies, CHASE 2017*, 2017. doi: 10.1109/CHASE.2017.62.

[19] H. Peng and M. You, “The Health Care Fraud Detection Using the Pharmacopoeia Spectrum Tree and Neural Network Analytic Contribution Hierarchy Process,” in *2016 IEEE Trustcom/BigDataSE/ISPA*, IEEE, Aug. 2016, pp. 2006–2011. doi: 10.1109/TrustCom 2016.0306.

[20] J. Rampal, P. Singh, R. Kaur, and K. Singh, “An Ensemble Model to predict Health Insurance Premium using Machine Learning,” *Journal-Dogorangsang.in*, vol. 10, no. 8, pp. 191–199, 2020.

[21] X. Zhao, Y. Zhang, S. Xie, Q. Qin, S. Wu, and B. Luo, “Outlier Detection Based on Residual Histogram Preference for Geometric Multi-Model Fitting,” *Sensors*, vol. 20, no. 11, 2020, doi: 10.3390/s20113037.

[22] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Front. Neurorobot.*, 2013, doi: 10.3389/fnbot 2013.00021.

[23] A. Stazio, J. G. Victores, D. Estevez, and C. Balaguer, “A Study on Machine Vision Techniques for the Inspection of Health Personnel’s Protective Suits for the Treatment of Patients in Extreme Isolation,” *Electronics*, vol. 8, no. 7, 2019, doi: 10.3390/electronics8070743.

[24] M. A. Morid, K. Kawamoto, T. Ault, J. Dorius, and S. Abdelrahman, “Supervised Learning Methods for Predicting Healthcare Costs: Systematic Literature Review and Empirical Evaluation,” *AMIA ... Annu. Symp. proceedings. AMIA Symp.*, 2017.

[25] C. Yang, C. Delcher, E. Shenkman, and S. Ranka, “Machine learning approaches for predicting high cost high need patient expenditures in health care,” *Biomed. Eng. Online*, vol. 17, no. S1, pp. 81–118, Nov. 2018, doi: 10.1186/s12938-018-0568-3.

[26] Rajiv, C., Mukund Sai, V. T., Venkataswamy Naidu, G., Sriram, P., & Mitra, P. (2022). Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic. *J Contemp Edu Theo Artific Intel: JCETAI/102*.

[27] Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. (2022). Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks. *J Contemp Edu Theo Artific Intel: JCETAI/101*.

[28] Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2020). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

[29] Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

[30] Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2022). Designing an Intelligent Cybersecurity Intrusion Identify Framework Using Advanced Machine Learning Models in Cloud Computing. *Universal Library of Engineering Technology*, (Issue).

[31] Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2022). Leveraging Artificial Intelligence Algorithms for Risk Prediction in Life Insurance Service Industry. *Available at SSRN 5459694*.

[32] Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(3), 70-80.

[33] Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies. *International Research Journal of Economics and Management Studies IRJEMS*, *1*(2).

[34] Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2022). Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations. *International Journal of Emerging Research in Engineering and Technology*, *3*(3), 99-107.

[35] Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Polu, A. R. (2022). Revolutionizing Marketing Analytics: A Data-Driven Machine Learning Framework for Churn Prediction. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(2), 112-121.

[36] Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. BLOCKCHAIN TECHNOLOGY AS A TOOL FOR CYBERSECURITY: STRENGTHS, WEAKNESSES, AND POTENTIAL APPLICATIONS.

[37] Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2022). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

[38] Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

[39] HK, K. (2020). Design of Efficient FSM Based 3D Network on Chip Architecture. *INTERNATIONAL JOURNAL OF ENGINEERING*, *68*(10), 67-73.

[40] Krutthika, H. K. (2019, October). Modeling of Data Delivery Modes of Next Generation SOC-NOC Router. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

[41] Ajay, S., Satya Sai Krishna Mohan G, Rao, S. S., Shaunak, S. B., Krutthika, H. K., Ananda, Y. R., & Jose, J. (2018). Source Hotspot Management in a Mesh Network on Chip. In *VDAT* (pp. 619-630).

[42] Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv preprint arXiv:1001.3781*.

[43] Gopalakrishnan Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv e-prints*, arXiv-1001.

[44] Krutthika H. K. & A.R. Aswatha. (2021). Implementation and analysis of congestion prevention and fault tolerance in network on chip. *Journal of Tianjin University Science and Technology, 54*(11), 213–231. https://doi.org/10.5281/zenodo.5746712

[45] Krutthika H. K. & A.R. Aswatha. (2020). FPGA-based design and architecture of network-on-chip router for efficient data propagation. *IIOAB Journal, 11*(S2), 7–25.

[46] Krutthika H. K. & A.R. Aswatha (2020). Design of efficient FSM-based 3D network-on-chip architecture. *International Journal of Engineering Trends and Technology, 68*(10), 67–73. https://doi.org/10.14445/22315381/IJETT-V68I10P212

[47] Krutthika H. K. & Rajashekhara R. (2019). Network-on-chip: A survey on router design and algorithms. *International Journal of Recent Technology and Engineering, 7*(6), 1687–1691. https://doi.org/10.35940/ijrte.F2131.037619

[48] Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 55-65.

[49] Gangineni, V. N., Tyagadurgam, M. S. V., Chalasani, R., Bhumireddy, J. R., & Penmetsa, M. (2021). Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management. *International Journal of AI, BigData, Computational and Management Studies*, *2*, 10-63282.

[50] Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. (2021). An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 26-34.

[51] Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. (2021). Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(3), 39-46.

[52] Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2021). Enhancing IoT (Internet of Things) Security Through Intelligent Intrusion Detection Using ML Models. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 27-36.

[53] Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2021). Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis. *International Journal of Emerging Research in Engineering and Technology*, *2*(3), 61-70.

[54] Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2021). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(2), 26-34.

[55] Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2021). Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(4), 54-61.

[56] Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. (2021). Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques. *Available at SSRN 5297803*.

[57] Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. *International Journal of Scientific Research and Modern Technology (IJSRMT) Volume*, *1*.

[58] Dinesh, K. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. *International Advanced Research Journal in Science, Engineering and Technology*.

[59] Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, *2*(2), 55-62.

[60] Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, *1*(1), 150-157.

[61] Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, *1*(1), 1-13.

[62] Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.