Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Predictive Analytics in Healthcare: A Hybrid Model for Efficient Medical Cost Estimation for Personalized Insurance Pricing

Dr. R. Bhargavi

bhargavi.r@vit.ac.in

Vellore Institute of Technology University

Dr. Subramanian Arumugam

Vellore Institute of Technology University

Article

Keywords: Insurance, Medical cost, Healthcare, Linear Regression, Gaussian, Deep Belief Network

Posted Date: August 30th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4697357/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

PREDICTIVE ANALYTICS IN HEALTHCARE: A HYBRID MODEL FOR EFFICIENT MEDICAL COST ESTIMATION FOR PERSONALIZED INSURANCE PRICING

Dr. R. Bhargavi [1] and Dr. Subramanian Arumugam [2]

School of Computing Science and Engineering, Vellore Institute of Technology (VIT), Chennai, Tamil Nadu, India [1&2]

E-Mail: bhargavi.r@vit.ac.in [1] and subramanian.a2014phd1136@vit.ac.in [2]

Corresponding Author: Dr. R. Bhargavi (E-Mail: bhargavi.r@vit.ac.in) [1]

ABSTRACT

Predictive analytics in healthcare enables personalized insurance pricing by analyzing diverse datasets, including patient histories and treatment outcomes. In the recent years, Machine Learning (ML) models and statistical approaches are used to enhance predictive analytics in healthcare by accurately identifying individual risk factors and forecasting medical costs from complex datasets, enabling more personalized insurance pricing. However, these models are not adept in capturing the intricate relationships between linear and non-linear patterns within healthcare data. This limitation can restrict their effectiveness in personalizing insurance pricing to the fullest extent, underscoring the need for innovative solutions that can bridge this gap. This paper presents the Linear Regression-Gaussian Deep Belief Network (LR-GDBN), a hybrid predictive model developed to improve healthcare analytics by accurately forecasting medical costs. This model effectively captures linear and non-linear patterns in healthcare data, surpassing traditional methods. LR-GDBN is evaluated with a standard dataset, and it demonstrates a low prediction error of 0.4926 surpassing conventional statistical and machine learning models. Unlike the state-of-the-art models capable of handling only non-linear health data in insurance pricing, this model featuring ability to handle linear and non-linear data is unique of its kind in the domain of healthcare underwriting.

Keywords: Insurance, Medical cost, Healthcare, Linear Regression, Gaussian, Deep Belief Network

1. INTRODUCTION

Predictive analysis in healthcare represents a pivotal shift towards data-driven decision-making, employing artificial intelligence (AI), ML, and sophisticated data mining techniques [1]. This approach processes extensive datasets from historical and real-time sources to uncover patterns that predict future health outcomes, enabling healthcare providers to intervene earlier, tailor treatments to individual patients, and optimize the allocation of resources more effectively. By foreseeing health events with a significant degree of confidence, predictive analytics aims to improve patient care, reduce unnecessary costs, and enhance the overall efficiency of healthcare systems. Its implementation, however, necessitates careful consideration of data integrity, privacy safeguards, and ethical standards to ensure patient trust and compliance with regulatory frameworks [2].

Predictive analytics is integral to forecasting medical costs, enhancing patient care, and maximizing healthcare efficiency by mitigating fraud and optimizing resource allocation. It enables healthcare providers to analyze a wide range of data, including patient histories, treatment outcomes, and resource utilization, to identify trends and predict future costs [3]. This analytical approach supports proactive public health management, allowing for informed decision-making, operational improvements, and significant cost reductions. Through the effective application of predictive analytics, healthcare organizations achieve better expense management, ensuring the judicious use of resources and improved patient outcomes.

The global health insurance market is experiencing robust growth as depicted in Figure 1, primarily driven by escalating healthcare costs, increased disposable incomes, and advancements in chronic disease treatments. This growth trajectory is underscored by a surge in health insurance product offerings and strategic initiatives by market players. The adoption of health insurance policies is on the rise, propelled by growing awareness of their importance. Notably, private health insurance spending reached $1,289.8 billion in 2022, representing 29% of the total national health expenditures. Future projections indicate that the insured population will reach 89.8% by 2030, with the private health insurance sector expected to expand at a 7.6% compound annual

growth rate (CAGR) from 2021 to 2030 [4,5]. The Asia Pacific region emerges as a significant contributor to this market growth, driven by heightened health insurance awareness, rising healthcare costs, and the continuous development of new healthcare treatments and services.

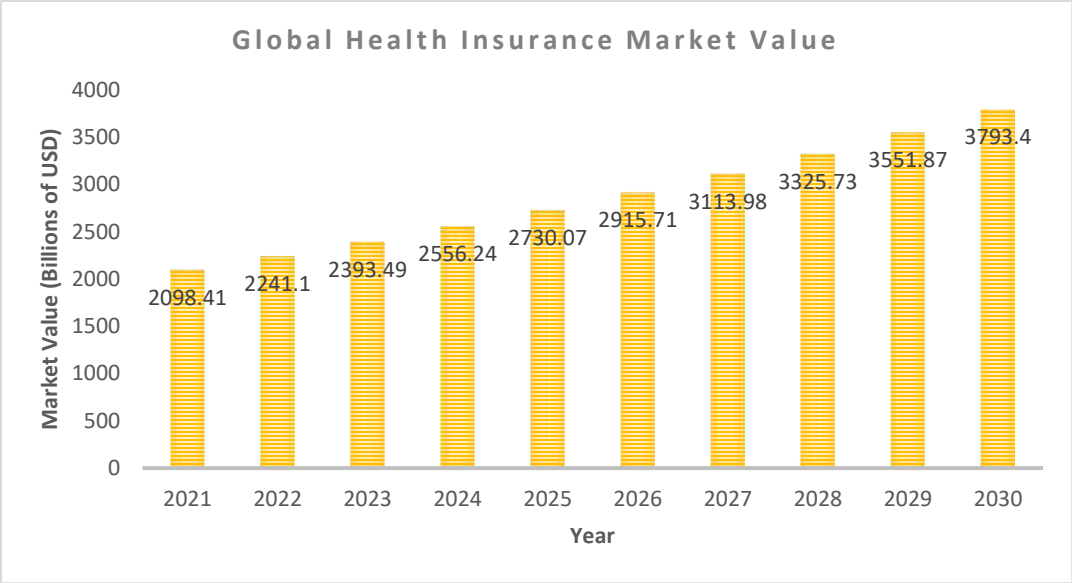


Figure 1. Global Health Insurance Market Trend

In the insurance sector, predictive analytics utilizes ML approacches to sift through vast datasets, enhancing the precision of medical cost predictions and underwriting processes. It facilitates the integration of varied data sources, allowing insurers to streamline claims processing, enhance data quality, and detect complex patterns [6,7]. This technology enables the creation of detailed customer profiles, optimization of pricing strategies, and identification of high-cost medical scenarios, ensuring more accurate premium settings.

Predictive analytics in healthcare primarily focuses on identifying high-risk patients to significantly enhance their care [8]. By leveraging ML algorithms and analyzing comprehensive datasets, healthcare insurers can pinpoint individuals at increased risk for chronic diseases or adverse health outcomes. This targeted approach allows for the customization of treatment plans based on a patient's medical history, demographics, and other critical factors, fostering early interventions that can prevent disease progression and improve overall health outcomes. In doing so, predictive analytics not only ensures more effective and personalized patient care but also contributes to greater efficiency and cost savings within the healthcare system.

Recent studies in the domain of health insurance cost prediction have leveraged various ML algorithms to enhance prediction accuracy and efficiency. Notably, research by Ramya and Deepa [9] compared the efficacy of Multinomial Logistic Regression (MLR) and Random Forest (RF) algorithms, emphasizing the importance of algorithm selection. Similarly, Choi et al. [10] developed a predictive model for high medical expenses using logistic regression, RF, and XGBoost, with XGBoost showing superior performance. The study by Iwendi et al. [11] demonstrated the use of a boosted Random Forest algorithm for COVID-19 patient health prediction, achieving significant accuracy. Furthermore, Chittilappilly et al. [12] explored the optimization of medical insurance prediction using a genetic algorithm and other ML techniques, highlighting the potential of automated systems.

The limitations of health insurance cost prediction models primarily stem from data quality issues, the complexity of healthcare data, and the models' interpretability challenges. Inaccuracies can arise from incomplete or biased datasets, while the high dimensionality and non-linear relationships in healthcare data complicate model training, risking overfitting or underfitting. Complex models like RF and XGBoost, despite their effectiveness, suffer from limited transparency, making it difficult for users to understand how predictions are made. Furthermore, these models require regular updates to remain relevant, posing additional resource demands.

Ethical concerns, including privacy and algorithmic bias, also present significant challenges, necessitating careful consideration to ensure equitable and responsible application.

Previous research in health insurance leveraging machine learning primarily utilized Linear Regression (LR) for various predictive tasks, including price prediction and expenditure estimation. However, LR models exhibit limitations, such as high variance with categorical variables, inadequacy in handling multiple classes, and difficulty in modeling non-linear relationships due to their linear assumptions. Given that many factors influencing insurance premiums—like age, gender, demographics, etc. exhibit non-linear relationships with premiums, there's a clear need for models capable of capturing these complex dynamics. To date, no hybrid models combining LR with deep learning to address both linear and non-linear relationships have been reported in the literature. This gap underscores the potential for developing such hybrid models to improve prediction accuracy in insurance premium pricing.

Integrating LR with Deep Belief Networks (DBNs) for health insurance premium prediction can offer a powerful solution by combining LR's proficiency in linear analysis with DBNs' capability to model complex, non-linear relationships inherent in insurance data. This approach improves accuracy by addressing both linear and non-linear factors like age, gender, and lifestyle, enabling precise, personalized premiums and better risk management. It represents a significant step forward in developing advanced insurance pricing models that accurately reflect individual risk profiles, optimizing product offerings and customer satisfaction. This study introduces the LR-DBN model, a novel hybrid approach designed to forecast individual future medical costs, thereby assisting insurance companies in determining personalized premiums. This innovative model synergizes the strengths of LR and DBN, offering a sophisticated tool for accurately predicting healthcare expenses. Through the integration of LR's linear analytical capabilities with DBN's proficiency in modeling complex, non-linear relationships, the LR-DBN model emerges as a cutting-edge solution for enhancing premium calculation processes. This advancement not only promises to refine the accuracy of insurance pricing but also paves the way for more customized insurance offerings, aligning premiums more closely with individual risk profiles.

The contributions of this research are as below.

1. The development of the LR-DBN hybrid model introduces a groundbreaking approach to predicting health insurance premiums, significantly improving accuracy by capturing both linear and non-linear data patterns through the novel integration of Linear Regression and Deep Belief Networks.
2. The use of Contrastive Divergence for training the GDBN enhances the model's efficiency and learning capability, enabling the hybrid model to achieve superior performance in health insurance cost prediction and offering scalability for broader applications.

The remainder of this paper is structured as follows: Section 2 offers a review of key works pertinent to this study. Section 3 outlines the dataset used and the methodologies applied in developing the proposed model. The architecture of the LR-GDBN model is detailed in Section 4. Section 5 presents the empirical results and insights derived from the research. Finally, the paper concludes in Section 6, providing suggestions for future research directions.

2. RELATED WORK

In the context of health insurance premium prediction, the literature reveals a progression from traditional statistical methods to sophisticated ML techniques to address the complexities inherent in predicting insurance costs. Historically, LR has been a foundational tool for predicting health insurance premiums, valued for its simplicity and interpretability. However, its limitations in capturing complex, non-linear relationships in data have led researchers to explore more advanced models. Recent studies have employed RF and neural networks to overcome the limitations of LR. For instance, Prakash et al. [13] utilized Random Forest Regression (RFR) with hyperparameter tuning for medical insurance premium prediction, achieving a notable accuracy of 98%, highlighting the potential of these models to enhance prediction accuracy by capturing complex patterns in the data. Similarly, Das et al. [14] developed a higher-order neural network approach to predict life insurance premiums during the pre- and post-Covid-19 era, demonstrating the superiority of advanced machine learning methods over traditional models in adapting to changing economic and social conditions.

Moreover, the integration of Explainable Artificial Intelligence (XAI) with ML models, as explored by Bora et al. [15], represents a significant advancement in making complex models more interpretable and trustworthy for stakeholders. This approach not only enhances the prediction accuracy but also builds confidence in the use of ML for critical financial decisions like insurance premium setting. Ahmad et al. [16] further emphasize the importance of adopting ML with an adaptive approach to predict insurance premiums, suggesting that such approaches can help insurance companies optimize their services with greater accuracy, thereby reducing financial losses.

LR has been a cornerstone in the field of predictive modeling due to its simplicity and interpretability, particularly in the domain of health insurance premium prediction. This statistical method models the relationship between a dependent variable, such as insurance premiums, and one or more independent variables, like age, gender, and health status, assuming a linear relationship between them. The application of LR in insurance pricing is grounded in its ability to provide a clear, understandable model that can be easily communicated to both stakeholders and customers. However, the evolving landscape of health insurance requires models that can accommodate the complexity and non-linearity of real-world data.

Recent scholarly efforts have explored the integration of LR with other predictive models to enhance the accuracy of health insurance premium predictions. For instance, the work by Jyothsna et al. [17] compared various regression models, including Multi-Linear Regression, and found that while Gradient Boosting outperformed with an accuracy of 87%, the inclusion of LR in the comparative analysis underscores its continued relevance in predictive modeling. This study exemplifies the ongoing exploration of combining traditional and advanced techniques to refine prediction accuracy.

Similarly, Ghosh and Gor [18] demonstrated the effectiveness of Multiple Linear Regression in predicting health insurance premiums within a blockchain technology framework. Their findings suggest that LR can yield superior results when applied to structured data, highlighting its adaptability and enduring value in the face of emerging technologies. Recent research reported in [19] advocates the application of LR among a suite of regression algorithms for medical insurance price prediction. Their approach, which assessed the performance of LR in conjunction with other models, reflects the comparative advantage and limitations of LR in a complex predictive landscape.

Moreover, the proposal of a Multiple Linear Regression model in [20] to predict health insurance charges based on demographic and health-related factors achieved notable accuracy. Their model's performance, with a Mean Absolute Percentage Error (MAPE) around 3% and an R^2 of 0.7615, indicates a significant improvement over traditional models, showcasing the potential of LR in enhancing decision-making processes in healthcare. These studies collectively highlight the ongoing relevance of LR in health insurance premium prediction, while also acknowledging its limitations in capturing complex patterns inherent in insurance data. The integration of LR with more sophisticated models represents a promising avenue for developing more accurate and personalized insurance pricing strategies. The paper by Alzoubi et al. [22] focuses on predicting medical insurance costs using machine learning regression techniques, comparing models like Gradient Boosting Regressor (GBR), AdaBoost Regressor, Lasso, and Elastic Net Regression on a US-based dataset. The study highlights boosting techniques outperforming regularization techniques in terms of R2 and minimized loss, aiming to assist organizations in creating more effective public-oriented medical insurance policies.

DBNs have emerged as a powerful tool in deep learning for modeling complex, non-linear relationships within large datasets, demonstrating significant versatility across various fields, including finance and insurance. These networks, characterized by their multi-layered architecture of stochastic variables, excel in uncovering hidden patterns in data, facilitating applications such as credit scoring, fraud detection, and risk assessment with greater accuracy than traditional ML models [23]. The strength of DBNs lies in their ability to handle high-dimensional data efficiently, a common challenge where traditional models falter, by leveraging their deep architecture to reduce dimensionality and extract relevant features through a combination of unsupervised pre-training and supervised fine-tuning [24]. This feature extraction capability, alongside the ability to capture the hierarchical structure of data, positions DBNs as particularly advantageous in fields burdened with complex data structures, such as image and speech recognition, natural language processing, and bioinformatics [25], underscoring their critical role in advancing both theoretical and applied aspects of machine learning and artificial intelligence.

Hybrid models leverage the strengths of various algorithms to tackle the complexities inherent in predicting insurance premiums. Notably, Yang et al. [26] developed a gradient tree-boosted Tweedie compound Poisson model, demonstrating superior performance in predicting insurance premiums by capturing complex interactions among predictors. This approach exemplifies the potential of hybrid models to address the limitations of traditional Generalized Linear Models (GLMs) by incorporating flexibility in capturing non-linear relationships and interactions. The evolution of hybrid models is also marked by the exploration of XAI techniques to enhance the interpretability of insurance premium predictions. Isa and Fernando [27] ventured the development of interpretable models that provide insights into the decision-making process, addressing the crucial need for transparency in AI-driven predictions.

Recently, in insurance analytics, the Tweedie distribution emerges as a pivotal model due to its adeptness in handling both zero claims and varying magnitudes of claim sizes, a common characteristic in insurance claim data. This distribution's flexibility is showcased across a spectrum of research efforts aimed at refining insurance premium calculations and ratemaking methodologies. Lian et al. [28] introduce an innovative approach by integrating the Tweedie compound Poisson model within a Reproducing Kernel Hilbert Space (RKHS), marking a significant advancement in modeling capabilities. This integration not only enhances the model's flexibility but also its precision in capturing complex patterns inherent in claim data, thereby facilitating more accurate insurance premium estimations. Building on the computational front, Sebyhed [29] explores the application of machine learning algorithms for insurance premium calculations based on claim models, including those governed by the Tweedie distribution. This study underscores the efficiency and predictive accuracy of machine learning techniques in navigating the complexities of claim data, offering a promising avenue for computational insurance analytics.

Further extending the application of machine learning in this domain, Huyghe et al. [30] present a novel technique involving boosting cost-complexity pruned trees on Tweedie responses, termed the ABT machine. This approach represents a leap forward in predictive modeling for insurance ratemaking, providing a robust framework for handling the intricacies of Tweedie-distributed data. Jeong [31] probes into the integration of credibility theory with the Tweedie distribution, proposing a multivariate semi-parametric credibility model. This contribution is particularly noteworthy for its sophisticated method of premium adjustment based on historical claim data, offering a nuanced perspective on risk assessment. Lastly, Gao [32] focuses on methodological advancements in fitting the Tweedie compound Poisson model to pure premium calculations using the Expectation-Maximization (EM) algorithm. This study highlights the EM algorithm's effectiveness in overcoming the computational challenges associated with maximum likelihood estimation for Tweedie models, thus enhancing the accuracy of pure premium estimates.

Despite the advancements brought by Tweedie distribution and machine learning techniques in insurance analytics, challenges such as model complexity, interpretability, data requirements, computational demands, and adaptability remain. These challenges underscore the necessity for innovative approaches, such as hybrid models that merge traditional statistical methods with advanced neural network architectures. This hybrid approach can balance interpretability with predictive accuracy, efficiently process sparse and skewed data and offer flexibility across various insurance contexts. The notable absence of hybrid modeling approaches in the domain of medical cost prediction for insurance premiums underscores a critical research gap that demands attention.

3. MATERIALS AND METHODS

This section outlines the dataset utilized for this study and the methodologies applied in implementing the proposed LR-DBN predictive model.

3.1 Dataset

The LR-DBN model employs the Kaggle medical cost personal dataset [33], encompassing 1338 records for its training and testing phases. These entries feature a mix of independent variables—Age, Sex, BMI, Children, Smoker, and Region—and a dependent variable, Charges. A sample subset of this dataset is given in Table 1, showcasing examples of these variables.

Age	Sex	BMI	Children	Smoker	Region	Charges
18	male	33.77	1	no	northeast	1725.52
19	female	27.9	0	yes	northwest	6884.92
37	female	27.74	3	no	southeast	7281.506
46	female	33.44	1	no	northwest	8240.59
28	female	33	3	no	southeast	4449.462
33	female	22.705	0	no	northeast	21984.47

Table 1. Kaggle Medical Cost Dataset

The attributes of this dataset are as below.
Age: Age of the beneficiary (18-64)
Sex: Gender of the beneficiary (male, female)
BMI: Body Mass Index (BMI) of the beneficiary (15.96-53.13)
Children: Number of children of the beneficiary (0-5)
Smoker: Whether an individual is a smoker or not (yes, no)
Region: Residential area of beneficiary in US (northeast, southeast, southwest, northwest)
Charges: Individual medical costs covered by health insurance ($1121.874 - $63770.43)
The dataset is divided into training and testing subsets following an 80:20 split. Additionally, the portion designated for training is further divided, adhering to a 70:30 split between training and validation, as depicted in Table 2.

No. of Training Samples	No. of Validation Samples	No. of Testing Samples
750	320	268

Table 2. Distribution of Dataset

3.2 Linear Regression

LR stands as a cornerstone statistical technique in field of data analysis, particularly revered for its simplicity and robustness in exploring and quantifying the relationships between variables. Central to the technique is the linear equation in (1), where y is the dependent variable to be predicted, x represents one or more independent variables that are presumed to influence y, m is the slope indicating the rate of change of y with respect to x, and b is the y-intercept. Figure 2 depicts this regression process. LR is widely used across various sectors, including finance and healthcare, for its predictive capabilities. It enables businesses to forecast sales trends, optimize inventory levels, and devise pricing strategies. Economists leverage LR to predict essential economic indicators such as inflation rates, employment levels, and GDP growth. In finance, it aids in forecasting asset performance, influencing investment decisions by estimating stock prices, bond yields, and currency exchange rates. LR's strength lies in its transparent, interpretable framework, allowing for the quantification of how specific variables affect outcomes. For example, in healthcare, LR can estimate patient costs by considering factors like age, treatment frequency, and health metrics, while in marketing, it can predict the effectiveness of advertising spend on return on investment.

$$y = mx + b \quad (1)$$

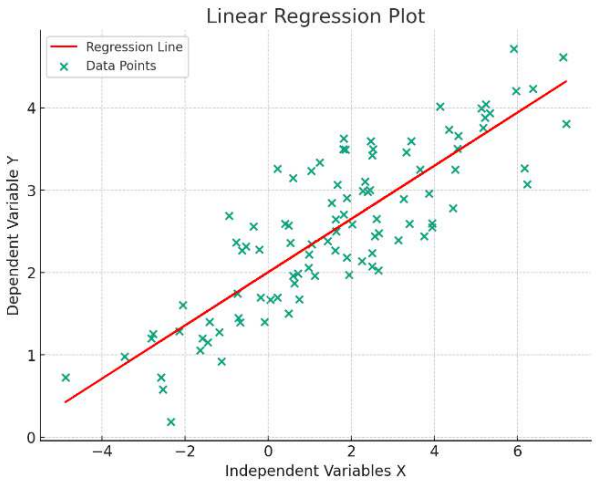


Figure 2. Linear Regression

Building an LR model entails a meticulous process of selecting relevant independent variables that are not only significant predictors but also devoid of multicollinearity. The chosen variables are then used to construct a model that captures the essence of the underlying data structure. The least squares method plays an integral role in this construction, as it identifies the optimal values of m and b that minimize the sum of squared residuals-the discrepancy between the observed values and those predicted by the model. This method ensures that the resulting model is the best fit for the data, providing a reliable basis for predictions. The accuracy and reliability of LR models, however, are contingent upon the quality and nature of the data at hand, as well as the appropriateness of the linear assumption in capturing the relationship between the variables.

In a dataset D containing pairs of observations $(x_1, y_1), \ldots, (x_n, y_n)$, the set $X = \{x_i, i = 1,2, \ldots n\}$ represents the independent variable, while and $Y = \{y_i, i = 1,2, \ldots n\}$ corresponds to the dependent variables. For n data points, the LR model is formulated as in (2) where y_i is the value of the dependent variable, β_0 is the model's intercept or bias, β_1 is the slope coefficient representing the effect of the independent variable x_i on y_i and ε denotes the random error component. The objective of the training process is to determine the best estimates for the parameters β_0 and β_1 that define the model.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon \tag{2}$$

The assessment of a LR model's accuracy typically involves the Mean Squared Error (MSE) metric, which is the average of the squares of the model's errors for all data points within the training set. This metric corresponds with the least squares method's objective to minimize the sum of the squared differences between observed values and the values predicted by the model. The line that achieves the smallest sum of squared residuals is termed the "best fit line." The mathematical representation of the sum of squared residuals, which the least squares method seeks to minimize, is given in equation (3).

$$S = \sum_{i=1}^{n} (y_i - (\beta_0 - \beta_i x_i))^2 \tag{3}$$

To minimize the sum of squared errors and find the best fit line for a LR model, one must calculate the partial derivatives of the sum of squared residuals function with respect to the intercept β_0 and the slope β_i. These derivatives are set to zero to solve for the coefficients that will yield the minimum error. The partial derivative with respect to β_0 is shown in (4), which simplifies to zero for the least squares function, while the partial derivative with respect to β_i is shown in (5). Typically, the partial derivative with respect to β_0 would yield an equation involving the sum of the errors (differences between observed and predicted y_i values), and the partial derivative with respect to β_i would involve the sum of the products of x_i values and their respective errors. Thus, the correct forms for the partial derivatives, set to zero for the purpose of finding the coefficients, are usually expressed as in (6).

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \ (y_i - (\beta_0 + \beta_1 x_i))^2 = 0 \quad (4)$$

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} \ (y_i - (\beta_0 + \beta_1 x_i))^2 = 0 \quad (5)$$

$$f(x) = \beta_0 + \beta_1 x \quad (6)$$

Ultimately, LR's simplicity in model formulation and interpretation, coupled with its predictive prowess, makes it an indispensable tool for analysts seeking to unearth trends, make forecasts, and inform decision-making across various disciplines and industries.

3.3 Deep Belief Network

A DBN is a type of generative graphical model that identifies intricate patterns within datasets via a hierarchical stratification of latent variables. It resembles a layered neural network where the uppermost layer corresponds to the observable data, while the successive, deeper layers consist of latent variables that capture progressively more abstract representations of the data. The architecture of a DBN is outlined by a Directed Acyclic Graph (DAG), which facilitates multiple tiers of hidden variables, making it adept at handling extensive datasets through efficient optimization techniques. DBNs are constructed from a stack of simpler models such as Restricted Boltzmann Machines (RBMs) or Hopfield networks, which collectively form its complex structure. The configurations of a DBN and an RBM are graphically represented in Figures 3 and 4 respectively.

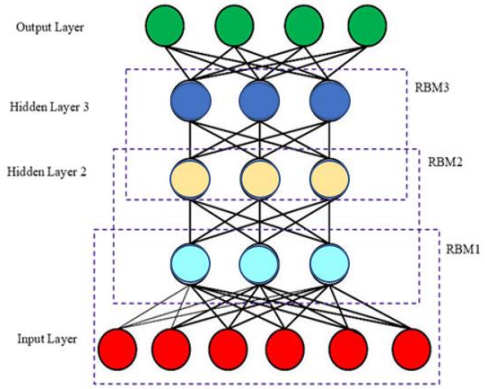


Figure 3. Deep Belief Network

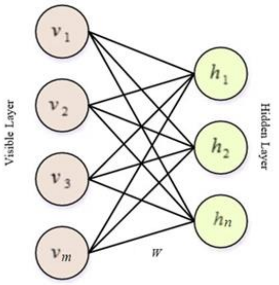


Figure 4. Restricted Boltzmann Machine

An RBM is a type of neural network that functions as an energy-based model with a bipartite architecture, including an input or Gaussian visible layer and a hidden layer. The Gaussian visible layer units represent observable variables, while the hidden layer encapsulates latent variables. The system's states are

evaluated through an energy function, and their probabilistic distribution is modeled using a Boltzmann distribution, anchored by this energy function. In the context of a DBN, RBMs are the building blocks. A DBN is constructed by sequentially stacking RBMs, where the hidden layer of one serves as the Gaussian visible layer for the next. This structure defines a DBN's learning process, which encompasses two phases: pretraining and fine-tuning. Pretraining engages in unsupervised learning to capture the underlying structure of the data, while fine-tuning employs supervised learning to refine the model parameters. The training of stacked RBMs in a DBN is typically executed through Contrastive Divergence (CD), an efficient algorithm for unsupervised learning. This step aims to find a parameter set that allows the model to accurately reconstruct input data. After establishing these initial parameters, the DBN undergoes fine-tuning, adjusting the weights in a supervised manner to optimize the network's predictive performance.

A Restricted Boltzmann Machine (RBM) is organized into two layers: a Gaussian visible layer denoted by $v = (v_1, v_2, \dots, v_m)$ and a hidden layer indicated by $h = (h_1, h_2, \dots, h_n)$. A more general approach to incorporating Gaussian distributions in RBMs, especially for the visible layer, is to modify the energy function to include the Gaussian distribution's characteristics. For a Gaussian RBM, where the visible units are assumed to follow a Gaussian distribution, the energy function (7) and the log probability (8) could be reformulated as follows, assuming a standard deviation of 1 for simplicity:

$$E(v,h) = \sum_{i=1}^{m} \frac{(v_i - b_i)^2}{2} - \sum_{i=1}^{m} \sum_{j=1}^{n} v_i w_{ij} h_j - \sum_{j=1}^{n} c_j h_j \quad (7)$$

$$\log p(v|\theta) = -E(v,h) - \log Z(\theta) \quad (8)$$

Here, b_i and c_j are the biases for the visible and hidden units, respectively, and w_{ij} represents the weights between visible unit i and hidden unit j. $Z(\theta)$ is the partition function, which is a normalization factor ensuring that the probabilities sum up to 1. This formulation reflects the Gaussian nature of the visible units by including a term $\frac{(v_i - b_i)^2}{2}$, which corresponds to the exponent part of the Gaussian distribution's probability density function, assuming a variance of 1 for simplicity.

The energy potential of an RBM is akin to the system's negative entropy, with the Gaussian visible layer serving as the input and the hidden layer as the output. The entropy of the dataset can be expressed through equation (9). This equation encapsulates the inherent uncertainty or randomness in the state of the hidden layer units, represented by h_i, within the RBM framework, where the energy associated with a node h_i is given as in (9). The expected entropy of the hidden layer units h_i in a discrete framework, where $h_i p(h_i))$ denotes the probability of each state is given in (10).

$$e_h(h_i) = h_i \log h_i - h_i - h_i \log (1 - h_i) \quad (9)$$

$$-\sum_i (h_i p(h_i)) = \int e_h(h_i) dh \quad (10)$$

The probability distribution of an RBM is given in equation (11), where σ is the parameter that governs the spread of the distribution. The likelihood of observing the Gaussian visible layer when the hidden layer is known is given in equation (12). To identify the optimal parameters θ of the RBM, one must maximize the Evidence Lower Bound (ELBO), as described in equation (13), where ψ_h and ψ_v are functions that map the states of hidden and Gaussian visible units to their respective probabilities, and $E(v,h)$ represents the energy of a state within the RBM. The ELBO maximization process involves adjusting θ to enhance the fit of the model to the observed data while also maintaining a balance with the model's complexity.

$$p(v,h) = e^{-\frac{E(v,h)}{\sigma^2}} = \prod_{i=1}^{m} [\psi_h(h_i)\psi_v(v_j) + \psi_h(1 - h_i)\psi_v(1 - v_i)] \quad (11)$$

$$p(h|v) = \frac{e^{-\frac{E(v,h)}{\sigma^2}}}{\sum_{h'} e^{-\frac{E(v,h')}{\sigma^2}}} \quad (12)$$

$$ELBO(\theta) = p(h \mid v) \log \left(\frac{p(h|v)}{p(h)}\right) + \int p(v \mid h) \log \left(\frac{p(v|h)}{p(v)}\right) dh \quad (13)$$

In the RHS of (13), the initial term represents the reconstruction likelihood, which pertains to the probability of reconstructing the Gaussian visible layer given the hidden layer's configuration. The latter term denotes the model's complexity, reflecting the entropy of the distribution over the Gaussian visible layer. For a DBN, an effective RBM is achieved by optimizing the ELBO, which involves enhancing the model's reconstruction accuracy while managing its complexity through entropy.

4. PROPOSED LR-GDBN MODEL FOR MEDICAL COST PREDICTION

This section presents the architecture of the proposed LR-GDBN framework for medical cost prediction. Figure 5 depicts the components of this framework. A detailed description of the components and workflow of the LR-GDBN framework is as below.

1. At the base, the LR model predicts the target outcome and calculates the prediction errors. These errors are then fed into the GDBN as input.
2. The GDBN is composed of several layers of interconnected nodes, each representing a different level of data abstraction. The initial layer of the GDBN receives the input from the LR model. As the information flows upward, each subsequent layer, structured as a RBM, processes the input to learn complex patterns and relationships within the data.
3. The training of these RBMs involves a technique that allows the network to model the data distribution effectively. This layered approach enables the LR-GDBN to capture both the direct relationships learned by the LR model and the more intricate, non-linear relationships discovered by the GDBN, leading to a comprehensive model that can make accurate predictions.
4. The final output of the framework is the predicted medical costs. By leveraging the feature extraction capabilities of the GDBN and the predictive modeling of LR, the LR-GDBN framework aims to provide accurate and reliable cost forecasts.

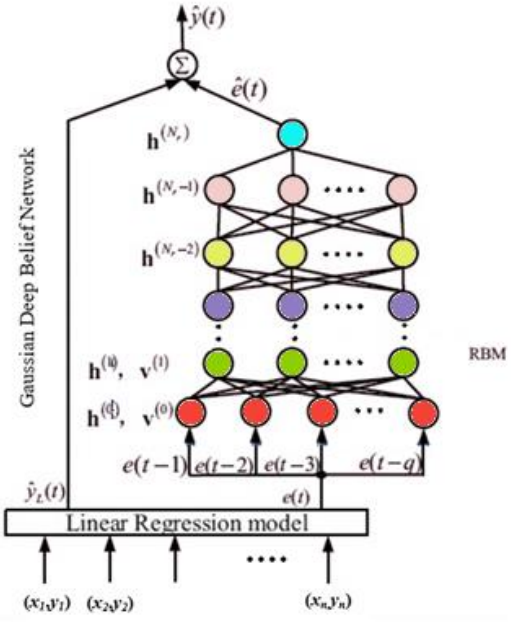


Figure 5. LR-GDBN Framework

The LR-GDBN framework embodies a synergistic approach that combines deep learning's ability to handle high-dimensional data with linear regression's straightforward interpretability. This makes it particularly well-suited for applications in healthcare analytics, where both complexity and transparency are paramount.

Initially, the LR model undergoes training with a designated training dataset, during which a cost function is formulated and subsequently minimized. This cost function quantifies the discrepancy between the predicted values by the model and the actual medical charges. It is mathematically expressed in equation (14), where y_i denotes the actual charges, $\hat{y_i}$ signifies the charges predicted by the model, and m represents the total number of data points within the training dataset.

$$e(y, \hat{y}) = \frac{1}{m}\sum_1^m \ (y_i - \hat{y_i})^2 \tag{14}$$

The cost function, due to its complexity, is non-linear, highlighting that a straightforward linear connection between the dependent and independent variables cannot be established. To address this non-linearity, the GDBN is structured to adapt to the non-linear aspect $e(y, \hat{y})$, as in equation (15). Here, $f(\cdot)$ denotes the function of the inner layers that needs to be inferred, while $g(\cdot)$ symbolizes the function of the feed-forward hidden layers. The parameters *W* and *b* are indicative of the weights and biases associated with these layers, respectively.

$$GDBN(W, b) = f(g(f(W, b))) \tag{15}$$

The GDBN undergoes training through the CD method. This technique, broadly applicable for training DBNs, focuses on refining the cost function utilizing the backpropagation algorithm. The steps involved in the CD based training method are outlined as follows.

Step 1: Initialize the weights w_{ij} for each hidden unit h_j as w_{ij}=0

Step 2: Initialize the "reconstruction error" $E_t = 0$ and "contrastive error" $E_{ct} = 0$.

Step 3: Sample a minibatch $\{y_i\}_{i=1}^m$from training dataset

Step 4: For each hidden unit h_j

a. Compute the forward probability $\{p_{y_i}\}_{i=1}^m = \{\frac{exp(\sum_{j'=1}^N w_{ij'}y_i)}{exp(\sum_{i'n=1}^m w_{ij'}y_{i'})}\}_{i=1}^m$

b. Compute the contrastive probability $\{p_j\}_{j=1}^N = \{\frac{exp(\sum_{j'=1}^N w_{ij'}y_i)}{exp(\sum_{j'=1}^N w_{ij'}g(h_j))}\}_{j=1}^N$

c. Update the weights $w_{ij} = w_{ij} + \frac{\gamma E_{ct}}{m}(y_i - g(h_j))p_{y_i}$where $g(h_j)$ is the activation function for the hidden unit $h_j$and γ is the self adaptive parameter

Step 5: Set $E_t = E_t + m\frac{\gamma E_{ct}}{m}(y_i - g(h_j))p_j$

Step 6: Repeat the above steps T times or until E_t converges

Step 7: After T iterations, the h_j 's are trained using the new weights

Step 8: Repeat step 1 to step 7 for each RBM unit

This algorithm outlines the training process of a GDBN using the CD method, specifically tailored for optimizing the weights of the hidden units to minimize reconstruction and contrastive errors. By following these steps, the GDBN, through the CD method, efficiently trains to capture the underlying statistical structure of the dataset, enabling it to predict or reconstruct the input data with high fidelity.

5. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental findings and interpretations of this study. It engages in a comprehensive discussion on the broader implications of these results, highlighting their relevance and potential impact on the field of medical cost prediction in personalized insurance pricing.

5.1Experimental Setup

Initially, the medical cost dataset undergoes several preprocessing steps to ensure its suitability for training the LR-GDBN. Continuous variables such as age and BMI are normalized to have a common scale, while categorical variables like sex and smoker status are transformed through encoding methods. This standardization facilitates the effective learning of patterns without bias towards any particular scale of data. The model's experimental validation is anchored in a robust high-performance computing infrastructure, equipped with advanced hardware specifications. This includes an array of Intel Xeon Gold 6248R CPUs, operating at a base frequency of 3.0 GHz and featuring 24 cores per processor, which synergistically propel the computational

throughput. The system is complemented by 256GB of DDR4 ECC RAM, ensuring rapid access and manipulation of large datasets. The computational nodes are interconnected through a 40 Gbps InfiniBand network, promoting accelerated data transfer rates essential for synchronous model training and data handling tasks. Python 3.8 and TensorFlow 2.4 provide the necessary software environment for effective model training and evaluation. This setup ensures the precise tuning of model parameters and facilitated the evaluation of the LR-GDBN's capability to predict medical costs.

Grid search is a comprehensive hyperparameter optimization technique that methodically searches through a specified subset of hyperparameters to find the combination that yields the best model performance. It operates by defining a grid of hyperparameter values and evaluating every possible combination of these values across the grid. In this research, this process involves training the LR-GDBN model multiple times with different hyperparameter sets and measuring performance with the MSE. The LR-GDBN model's hyperparameters were meticulously optimized via grid search to achieve an optimal mix of accuracy, complexity, and efficiency. A learning rate of 0.001 was selected to facilitate stable convergence, complemented by 100 epochs to ensure thorough learning. The choice of a batch size of 32 strikes a balance between computational efficiency and the advantages of gradient descent. Structured with three layers (256, 128, 64 units), the model adeptly captures complex data features, utilizing ReLU activation to avoid gradient vanishing problems. To combat overfitting, a dropout rate of 0.5 and L2 regularization at a rate of 0.01 were implemented. Furthermore, the training of the RBM component is refined through a single step of Contrastive Divergence. Additionally, an early stopping criterion, based on validation loss, ensures training halts when no improvement is observed over a defined number of epochs, further enhancing the model's robustness and generalizability for the accurate prediction of health insurance premiums. Table 3 lists the hyperparameters used in tuning the model and their optimal values.

Hyperparameter	**Value**
Learning Rate (LR)	0.001
Number of Epochs	100
Batch Size	32
Number of Layers in GDBN	3
Units per Layer in GDBN	[256, 128, 64]
Activation Function	ReLU
Dropout Rate	0.5
Regularization Type	L2
Regularization Rate	0.01
Contrastive Divergence Steps (CD)	1

Table 3. LR-GDBN Hyperparameters

5.2 Exploratory Data Analysis

Exploratory Data Analysis (EDA) was performed on the dataset intended for the LR-GDBN model, employing visualization techniques to understand the data's characteristics. EDA emphasizes exploring data and forming hypotheses, focusing on validating assumptions for hypothesis testing and model fitting, rather than initial data analysis. The aim was to uncover insights into the data's quality and structure as depicted in Figure 6.

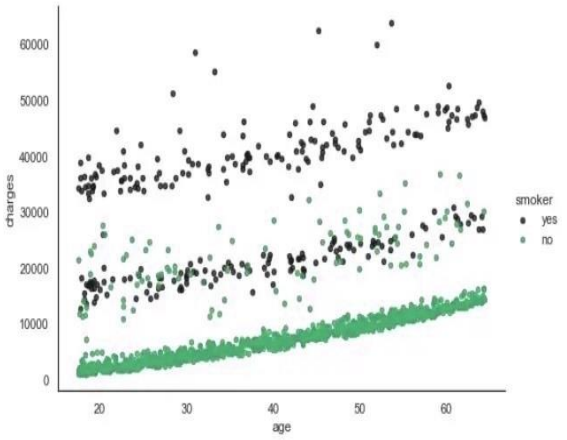
Figure 6a. illustrates the relationship between age and insurance charges, categorizing beneficiaries based on smoking status. The visualization reveals that smokers consistently incur higher premiums across all age groups. Additionally, the majority of beneficiaries with charges exceeding 40k are identified as smokers. The scatter plot in Figure 6b. displays insurance charges against BMI, distinguishing between smokers and non-smokers. It indicates that smokers generally face higher charges across all BMI levels, with the most substantial charges linked to smoking regardless of BMI. Non-smokers' charges remain comparatively lower and are less varied, suggesting smoking status may have a greater impact on insurance costs than BMI alone.

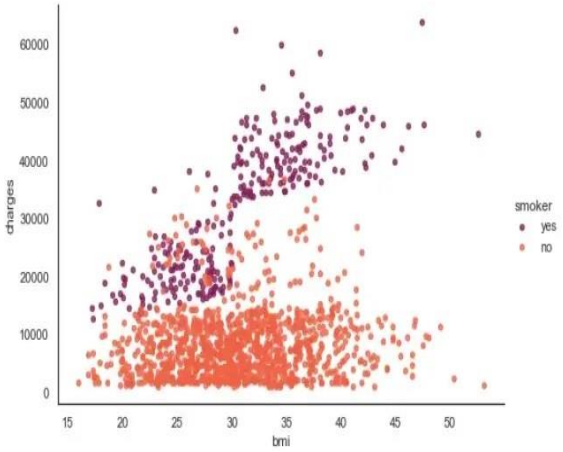
Similarly, Figure 6c. depicts a box plot comparison of insurance charges by gender, illustrating a higher median charge for males compared to females and a wider interquartile range, indicating greater variability among

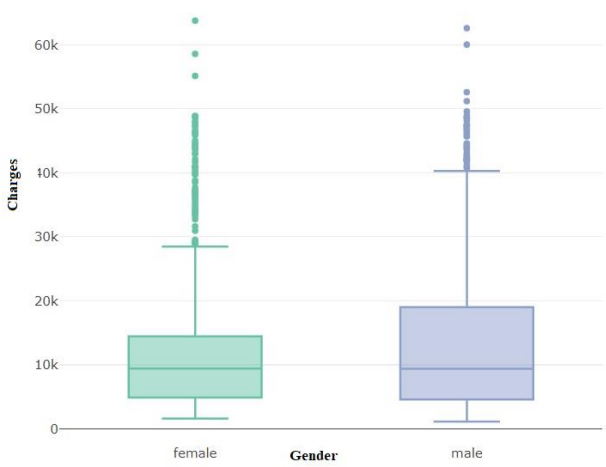
male charges. The presence of outliers in both categories suggests exceptions to general trends. This plot indicates that gender may play a role in the distribution of insurance charges, with males potentially incurring higher costs.

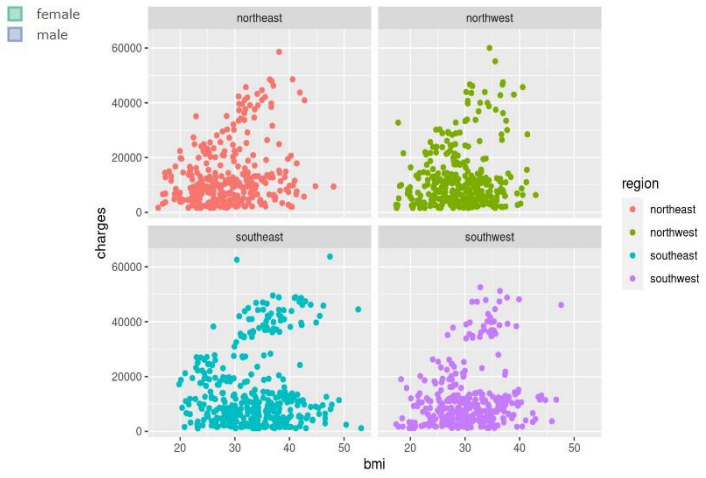
The scatter plots in Figure 6d. depict regional disparities in insurance charges relative to BMI. The Southeast region displays a wider range of charges and a higher concentration of data points at increased BMI levels, possibly indicating a correlation between higher BMI and higher charges in this region. The Northeast and North West regions show a more moderate spread of charges, with the North West cluster exhibiting lower charges across a similar BMI range. The Southwest region has the lowest range of charges and a more compact distribution of BMI values. This could infer that regional factors may influence the cost of insurance, with the Southeast region potentially facing higher healthcare expenses. Figure 6.e shows that insurance charges do not increase with the number of children; in fact, the median charges decrease for households with four and five children. The presence of outliers across all categories indicates that high insurance costs can occur regardless of family size. The data suggests a weaker correlation between the number of children and higher insurance charges, particularly for families with more than three children.

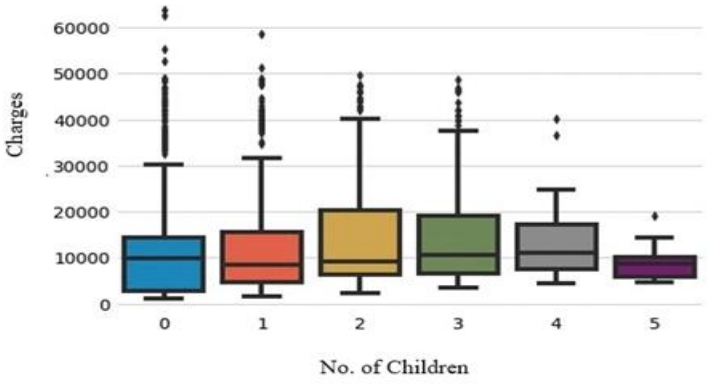
In addition to the above, the correlation matrix heatmap in Figure 6f. succinctly illustrates the relationships between demographics, habits, and insurance charges. Age shows a moderate positive correlation with charges, while BMI has a less pronounced but still positive relationship. The number of children bears little to no correlation with charges. Notably, a smoker's status is strongly positively correlated with charges, indicating that smokers tend to have higher insurance costs, a trend not observed with gender, which shows an insignificant correlation with charges. The matrix also reflects perfect negative correlations between dichotomous variables such as sex and smoker status, which is expected.











	age	bmi	children	charges	sex_female	sex_male	smoker_no	smoker_yes
age	1.000000	0.109272	0.042469	0.299008	0.020856	-0.020856	0.025019	-0.025019
bmi	0.109272	1.000000	0.012759	0.198341	-0.046371	0.046371	-0.003750	0.003750
children	0.042469	0.012759	1.000000	0.067998	-0.017163	0.017163	-0.007673	0.007673
charges	0.299008	0.198341	0.067998	1.000000	-0.057292	0.057292	-0.787251	0.787251
sex_female	0.020856	-0.046371	-0.017163	-0.057292	1.000000	-1.000000	0.076185	-0.076185
sex_male	-0.020856	0.046371	0.017163	0.057292	-1.000000	1.000000	-0.076185	0.076185
smoker_no	0.025019	-0.003750	-0.007673	-0.787251	0.076185	-0.076185	1.000000	-1.000000
smoker_yes	-0.025019	0.003750	0.007673	0.787251	-0.076185	0.076185	-1.000000	1.000000

Figure 6. Exploratory Analysis a. *Age* vs *Charges* b. *BMI* vs *Charges* c. *gender* vs *Charges* d. *region* vs *Charges* e. *children* vs *Charges* f. Correlation heatmap

The most critical inference from this analysis is that smoking status is a key variable associated with higher insurance charges. Age and BMI are also factors but with less impact than smoking. The number of children and gender do not appear to have a significant effect on insurance charges. This analysis is essential for insurers in risk assessment and premium calculation, and for individuals to understand factors that might influence their insurance expenses.

5.3 Performance Evaluation

The LR component of the LR-GDBN model is first trained on a subset of the data to predict medical costs, after which the residuals are calculated. Figure 7. illustrates the comparison between actual and predicted insurance charges in the test data, with the LR prediction line included. While a correlation is evident between the predicted and actual values, the figure also highlights many instances where the model's predictions diverge from the actual charges. These residuals, indicative of the LR model's prediction errors, are subsequently used to train the GDBN portion of the model. The scatter plot demonstrates the foundational strength of the LR component within the LR-GDBN model, successfully capturing the linear trend between actual and predicted insurance charges. The clear alignment of data points around the regression line, especially at lower charge values, affirms the model's robust predictive capability in this range. The dispersion of points at higher charge levels indicates an opportunity for the GDBN component to enhance the model by addressing non-linear complexities. This complementary relationship between the LR and GDBN components suggests that the LR-GDBN model leverages the simplicity of linear regression for lower values while harnessing the deep learning power of GDBN to improve prediction accuracy across the spectrum of charges.

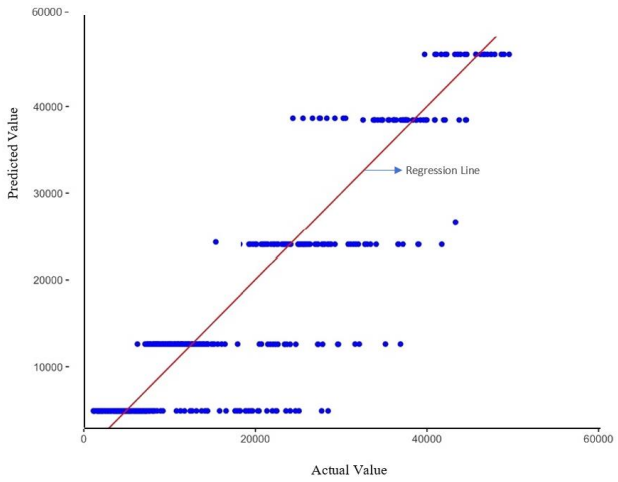


Figure 7. Medical cost prediction with LR Model

The performance of the LR-GDBN predictive model is assessed using three key metrics: MSE, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). The MSE is calculated as the average of the squares of the differences between actual and predicted values, formalized in equation (16). It quantifies the variance of the residuals. RMSE, given in equation (17), is the square root of the MSE and provides a measure of the standard deviation of the residuals. MAE, outlined in equation (18), is the average of the absolute differences between predicted and actual values, offering a linear score that reflects the average magnitude of the errors. Together, these metrics provide a comprehensive evaluation of the model's accuracy by highlighting different aspects of the prediction errors.

The MSE and the RMSE metrics equally penalize the magnitude of errors in predictions, while the MAE focuses on the absolute errors without weighting any deviations differently. The MAE offers a uniform measure of errors across various reference values, allowing for a balanced assessment of model performance, with lower values indicating superior predictive accuracy. The coefficient of determination R^2, is a statistical measure that quantifies the degree of variance in the data explained by the model. Calculated as in equation (19), R^2 involves the Error Sum of Squares (SSE) and the Total Sum of Squares (TSS), defined in equations (20-22). An R^2 value close to 1 suggests a high level of predictability of the dependent variable from the independent variables.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - y_i')^2 \quad (16)$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - y_i')^2} \quad (17)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - y_i'| \quad (18)$$

$$R^2 = 1 - \frac{SSE}{TSS} \quad (19)$$

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (20)$$

$$TSS = \sum_{i=1}^{n} (y_i - y)^2 \quad (21)$$

Adjusted R^2 (R^2_{adj}) modifies R^2 to account for the number of predictors in a model, ensuring a fair comparison, regardless of the number of independent variables. It is calculated as in equation (22) and provides a more accurate reflection of a model's predictive power when n observations and k variables are involved. R^2 and R^2_{adj} quantify the correlation between independent and dependent variables in LR. The R^2_{adj} compensates for model complexity and, if R^2 rises with more variables with a fall of R^2_{adj}, it indicates that new variables do not enhance the model meaningfully.

$$R^2_{adj} = 1 - \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \quad (22)$$

The performance of the LR-GDBN model is assessed using these metrics and its effectiveness is benchmarked against other regression models, including the basic LR, Ridge Regression (RR), Support Vector Regression (SVR), RFR, Long Short-Term Memory Regression (LSTM-R), and Convolutional Neural Network Regression (CNN-R), with the comparative results presented in Table 4.

Method /Model	MSE↓	RMSE↓	MAE↓	R^2↑	R^2_{adj}↑
LR	0.7719	0.8786	0.7510	0.6911	0.6711
RR	**0.6348**	**0.7967**	0.7026	0.8021	0.7789
SVR	0.7155	0.8459	0.7144	0.8164	0.7927
RFR	0.6612	0.8132	0.7264	0.8271	0.8031
LSTM-R	0.6881	0.8295	0.7386	0.8352	0.8109
CNN-R	0.7435	0.8622	**0.6911**	**0.8412**	**0.8168**
LR-GDBN	**0.4926**	**0.7018**	**0.5797**	**0.8757**	**0.8612**

Table 4. Performance Comparison

The LR-GDBN model excels in predicting medical costs, outperforming traditional models like LR and advanced models including RR, SVR, RFR, LSTM-R, and CNN-R, as shown in Table 4. It demonstrates the lowest MSE, RMSE, and MAE, indicating the highest accuracy and the smallest deviation from actual values. With the highest R^2 and R^2_{adj} values, LR-GDBN also demonstrates superior ability to explain variance in medical costs, significantly surpassing LR's lower performance metrics and even outdoing sophisticated models like LSTM-R and CNN-R.

While RR and SVR show improvements over LR in handling data complexity, and RFR, LSTM-R, and CNN-R present competitive capabilities, especially in capturing non-linear patterns, none match LR-GDBN's comprehensive performance. This analysis not only highlights the strengths of LR-GDBN in medical cost prediction but also underscores the importance of choosing a model that effectively addresses the data's complexity, where LR-GDBN stands out for its robust predictive power and efficiency in handling healthcare data intricacies. Figures 8 and 9 provide graphical depictions of the error metrics and squared R values, respectively, showcasing the comparative performance of different models in predicting medical costs.

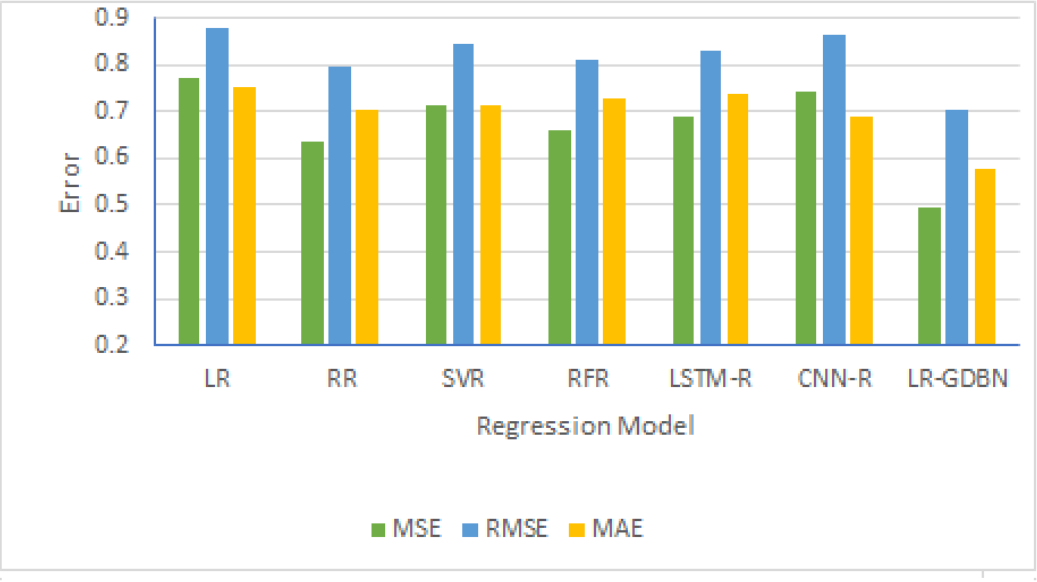


Figure 8. Prediction Errors of Regression Models

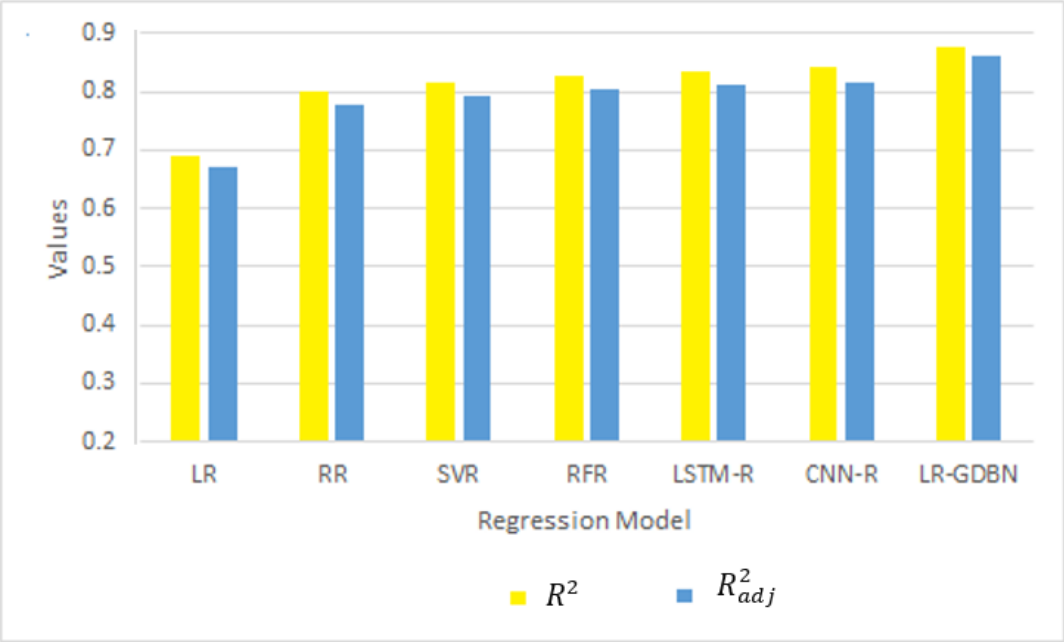


Figure 9. R2 and Adjusted R2 values of Regression Models

In addition to the above empirical evaluations, the proposed model is compared with most recent works in Table 5, with respect to important aspects.

Table 5. Comparison with State of the Art

Aspect	**LR-GDBN (Proposed)**	**Lian et al. (2023) [28]**	**Sebyhed (2023) [29]**	**Huyghe et al. (2023) [30]**	**Jeong (2023) [31]**	**Gao (2024) [32]**
Model Type	Hybrid Deep Learning	Statistical Model	Statistical Model	Machine Learning	Statistical Model	Statistical Model
Data Handling	Non-linear & Linear	Non-linear	Not Specified	Non-linear	Non-linear	Non-linear
Prediction Focus	Insurance Costs	Statistical Distribution	Insurance Premiums	Insurance Ratemaking	Credibility	Pure Premium
Algorithmic Approach	LR + GDBN	Kernel Methods	Claims-based	Boosted Trees	Semi-Parametric	EM Algorithm
Application Area	General Insurance	Technometrics	Machine Insurance	General Insurance	Multivariate Analysis	Insurance Economics
Model Characteristics	Integrative, Scalable, Flexible	High-dimensional Capability	Claim-focused	Pruning for Complexity, Efficient	Semi-Parametric Flexibility	Focused on Premium Estimation

The proposed LR-DBN model stands out for its integrative approach, combining the clear interpretability of LR with the sophisticated pattern recognition capabilities of a DBN. This hybridization not only enhances accuracy in capturing complex non-linear relationships but also retains the simplicity needed for understanding direct linear influences. As a result, the LR-DBN model is not only scalable and flexible but also well-suited for the multifaceted nature of insurance cost data. It offers a comprehensive solution that surpasses the limitations of single-method models, positioning it as a highly adaptable and effective tool for predicting medical costs in the insurance domain.

5.4Discussions

The experimental results reveal that the proposed hybrid LR-GDBN model leverages the strengths of a traditional LR model and the GDBN. The process begins with the LR model predicting insurance costs from the dataset, with the prediction errors (residuals) further used to refine the GDBN training. A Gaussian process within the GDBN ensures accurate mapping of these residuals to insurance costs. Performance comparisons reveal that the LR-GDBN model outperforms RR model by 14% in MSE and 9% in RMSE. Against the CNN-R, the LR-GDBN achieves superior performance with an 11% improvement in MAE, 18% in R^2, and 19% in R^2_{adj}. With the lowest MSE of 0.4926, RMSE of 0.7018, MAE of 0.5797, the highest R^2 of 0.8757, and $R^2_{adj}$of 0.8612, the LR-GDBN model stands out against both standard regression and CNN-based models. Its scalability and adaptability suggest potential applications in client classification, claim analysis, price optimization, and policy evaluation, underscoring its utility beyond the initial medical cost prediction task.

Training GDBNs with CD harnesses its ability to efficiently handle the probabilistic graphical models' parameters, enabling the network to capture the intricate correlations within the input data. This method is especially advantageous for modeling the non-linear and high-dimensional nature of insurance cost data, where traditional training methods might struggle. By employing CD, the GDBN quickly learns the distribution of the input data, effectively reducing the training time while maintaining or even improving the predictive accuracy. The implementation of CD in GDBN training distinguishes the LR-GDBN model from other regression models and deep learning approaches. It ensures that the GDBN component does not merely act as a passive recipient of the residuals from the LR model but actively refines these inputs to map accurately to insurance costs. This approach allows the GDBN to exploit the residuals' information content fully, making the hybrid model more adept at capturing complex patterns and relationships in the data that other models might miss.

Moreover, the use of CD in GDBN training contributes to the model's superior performance metrics, as evidenced by its significant improvements over RR and CNN-R in MSE, RMSE, MAE, R^2 and R^2_{adj}. This method's effectiveness is not only in enhancing the model's predictive performance but also in its potential applicability to a broader range of predictive tasks beyond medical cost prediction. The adaptability and scalability of the LR-GDBN model, powered by CD-trained GDBNs, offer promising avenues for its application in various domains such as client classification, claim analysis, price optimization, and policy evaluation, showcasing the versatility and robustness of this training approach.

The LR-GDBN model, while powerful, faces interpretability challenges due to the opaque nature of its deep learning component, making it difficult for stakeholders to understand the decision-making process. This can be mitigated by applying model-agnostic interpretability tools like Local Interpretable Model-agnostic Explanations (LIME) or SHapley Additive exPlanations (SHAP) to provide transparency. Additionally, the model's dependence on large datasets for training poses a constraint in data-scarce scenarios. Adopting transfer learning techniques, where a model is pre-trained on abundant data from a related domain and fine-tuned for the specific task, or employing data augmentation methods, can effectively reduce the need for extensive data, making the LR-GDBN more adaptable to varied insurance contexts with limited datasets.

6. CONCLUSION

This paper addresses the challenge of accurately predicting health insurance premiums, which is a complex problem due to the intricate and non-linear relationships inherent in medical cost data. The proposed LR-GDBN model stands as a novel solution, harnessing the strengths of Linear Regression and Gaussian Deep Belief Networks to capture both linear and non-linear patterns within the data. This hybrid approach has demonstrated superior predictive performance compared to traditional models, as evidenced by various performance metrics. The scope for extension of the LR-GDBN model is substantial. Its flexible architecture allows for scalability and adaptability to a wide range of insurance-related tasks, such as client classification, claim analysis, price optimization, and policy evaluation. The model’s ability to handle diverse data types and structures makes it a promising tool for broader applications in the actuarial field such as client classification, risk assessment, price optimization, and policy evaluation.

Data Availability

The datasets generated and/or analysed during the current study are available in Kaggle repository.
URL: https://www.kaggle.com/datasets/mirichoi0218/insurance

REFERENCES

1. Muniasamy, A., Tabassam, S., Hussain, M. A., Sultana, H., Muniasamy, V., & Bhatnagar, R. (2020). Deep learning for predictive analytics in healthcare. In *The International Conference on Advanced Machine Learning Technologies and Applications (AMLTA2019) 4* (pp. 32-42). Springer International Publishing.
2. Guo, C., & Chen, J. (2023). Big data analytics in healthcare. In *Knowledge Technology and Systems: Toward Establishing Knowledge Systems Science* (pp. 27-70). Singapore: Springer Nature Singapore.
3. Broekharst, D. S., van de Wetering, R., Ooms, W., Helms, R. W., & Roijakkers, N. (2023). Deploying predictive analytics to enhance patient agility and patient value in hospitals: A position paper and research proposal. *Healthcare Analytics*, *3*, 100141.
4. https://www.precedenceresearch.com/healthcare-insurance-market
5. https://www.cms.gov/data-research/statistics-trends-and-reports/national-health-expenditure-data/nhe-fact-sheet
6. Frees, E. W., & Gao, L. (2020). Predictive analytics and medical malpractice. *North American Actuarial Journal*, *24*(2), 211-227.
7. Oesterreich, T. D., Fitte, C., Behne, A., & Teuteberg, F. (2020). Understanding the role of predictive and prescriptive analytics in healthcare: a multi-stakeholder approach.
8. Parikh, R. B., Gdowski, A., Patt, D. A., Hertler, A., Mermel, C., & Bekelman, J. E. (2019). Using big data and predictive analytics to determine patient risk in oncology. *American Society of Clinical Oncology Educational Book*, *39*, e53-e58.
9. Ramya, D., & Deepa, J. (2022, October). Health Insurance Cost Prediction using Machine Learning Algorithms. In 2022 International Conference on Edge Computing and Applications (ICECAA) (pp. 1381-1384). IEEE.
10. Choi, Y., An, J., Ryu, S., & Kim, J. (2022). Development and Evaluation of Machine Learning-Based High-Cost Prediction Model Using Health Check-Up Data by the National Health Insurance Service of Korea. International Journal of Environmental Research and Public Health, 19(20), 13672.
11. Iwendi, C., Bashir, A. K., Peshkar, A., Sujatha, R., Chatterjee, J. M., Pasupuleti, S., ... & Jo, O. (2020). COVID-19 patient health prediction using boosted random forest algorithm. Frontiers in public health, 8, 357.
12. Chittilappilly, R. M., Suresh, S., & Shanmugam, S. (2023, May). A comparative analysis of optimizing medical insurance prediction using genetic algorithm and other machine learning algorithms. In 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI) (pp. 1-6). IEEE.
13. Prakash, V. S., Bushra, S. N., Subramanian, N., Indumathy, D., Mary, S. A., & Thiagarajan, R. (2022). Random forest regression with hyper parameter tuning for medical insurance premium prediction. *International Journal of Health Sciences*, *6*(S6), 7093-7101.
14. Das, S., Nayak, J., Nayak, S., & Dey, S. (2022). Prediction of Life Insurance Premium during Pre-and Post-Covid-19: A Higher-Order Neural Network Approach. *Journal of The Institution of Engineers (India): Series B*, *103*(5), 1747-1773.
15. Bora, A., Sah, R., Singh, A., Sharma, D., & Ranjan, R. K. (2022, October). Interpretation of machine learning models using xai-a study on health insurance dataset. In *2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)* (pp. 1-6). IEEE.
16. Ahmad, S., Agarwal, A., & Ansari, H. (2023, July). Prediction of Insurance Premium using Machine Learning with an Adaptive Approach. In *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)* (pp. 1-5). IEEE.

17. Jyothsna, C., Srinivas, K., Bhargavi, B., Sravanth, A. E., Kumar, A. T., & Kumar, J. S. (2022, May). Health Insurance Premium Prediction using XGboost Regressor. In *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)* (pp. 1645-1652). IEEE.
18. Ghosh, M., & Gor, R. (2022). Health Insurance Premium Prediction Using Blockchain Technology And Random Forest Regression Algorithm. *International Journal of Engineering Science Technologies*, *6*(3), 74–82.
19. Vijayalakshmi, V., Selvakumar, A., & Panimalar, K. (2023, January). Implementation of Medical Insurance Price Prediction System using Regression Algorithms. In *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)* (pp. 1529-1534). IEEE.
20. Kumar Sharma, D., & Sharma, A. (2020). Prediction of Health Insurance Emergency Using Multiple Linear Regression Technique. *Eur. J. Mol. Clin. Med*, *7*, 98-105.
21. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.
22. Alzoubi, H. M., Sahawneh, N., AlHamad, A. Q., Malik, U., Majid, A., & Atta, A. (2022, October). Analysis Of Cost Prediction In Medical Insurance Using Modern Regression Models. In *2022 International Conference on Cyber Resilience (ICCR)* (pp. 1-10). IEEE.
23. Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*.
24. Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009, June). Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning* (pp. 609-616).
25. Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4*(5), 5947.
26. Yang, Y., Qian, W., & Zou, H. (2018). Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models. *Journal of Business & Economic Statistics*, *36*(3), 456-470.
27. Isa, U. A., & Fernando, A. (2022). Explainable AI and Interpretable Model for Insurance Premium Prediction.
28. Lian, Y., Yang, A. Y., Wang, B., Shi, P., & Platt, R. W. (2023). A Tweedie Compound Poisson Model in Reproducing Kernel Hilbert Space. *Technometrics*, *65*(2), 281-295.
29. Sebyhed, H. (2023). MACHINE INSURANCE PREMIUM CALCULATIONS BASED ON CLAIM MODELS.
30. Huyghe, J., Trufin, J., & Denuit, M. (2023). Boosting cost-complexity pruned trees on Tweedie responses: the ABT machine for insurance ratemaking. *Scandinavian Actuarial Journal*, 1-23.
31. Jeong, H. (2023). Tweedie Multivariate Semi-Parametric Credibility.
32. Gao, G. (2024). Fitting Tweedie's compound Poisson model to pure premium with the EM algorithm. *Insurance: Mathematics and Economics*, *114*, 29-42.
33. https://www.kaggle.com/datasets/mirichoi0218/insurance

Acknowledgements
We want to express our gratitude to all the study participants and the individuals involved in the research. This study was funded by Vellore Institute of Technology, Tamil Nadu, India

Authors and Affiliations
School of Computing Science and Engineering, Vellore Institute of Technology (VIT), Tamil Nadu, India.
Dr. R. Bhargavi and Dr. Subramanian Arumugam

Author contributions
Dr. Subramanian Arumugam had done the research and written the manuscript, and the validation was done by Dr. R. Bhargavi. She monitored the research and gave her insights during the entire process.

Corresponding author
Correspondence to Dr. R. Bhargavi (*bhargavi.r@vit.ac.in*)

Competing interests
The authors declare no competing interests.

Funding
Open access funding provided by Vellore Institute of Technology (VIT), Tamil Nadu, India.