International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

Health Insurance Claim Prediction and Risk Assessment Using an AI-Based Machine Learning Framework

Dr K. SREEENIVASULU[1], E. SREE RAMULU[2]

[1]Professor, Dept. of C.S.E, Anantha Lakshmi Institute of Technology and sciences Anantapur – 515721

[2]PG Scholar, Dept. of C.S.E, Anantha Lakshmi Institute of Technology and sciences Anantapur- 515721

Abstract: The rising prevalence of health issues and increasing medical expenditures have made health insurance an essential component of modern society. Insurance providers such as LIC, ICICI, HDFC ERGO, and Star Health offer financial support to individuals by covering medical expenses; however, accurately estimating claim amounts remains a challenging task. This study utilizes a comprehensive dataset comprising generalized, hospitalization, and claim-related data, where each record reflects an individual's insurance charges along with demographic and health attributes. Machine learning (ML) techniques are employed to analyse patterns in hospitalization costs and insurance payments, enabling effective grouping and prediction of claim amounts. In the contemporary insurance domain, data-driven approaches play a crucial role in risk assessment and financial planning. ML models are capable of extracting complex, non-linear relationships from historical data, including patient demographics, treatment details, and past claims. This study proposes a computational intelligence-based framework using algorithms such as Random Forest and Support Vector Machines (SVM) to predict health insurance expenses. Experimental results demonstrate that Random Forest achieves the highest prediction accuracy of 90%, outperforming SVM with 83%. The comparative analysis highlights the effectiveness of ensemble learning methods in improving predictive performance. The proposed approach supports better decision-making, risk management, and resource allocation in the healthcare insurance sector.

Key Words: Health Insurance Prediction, Machine Learning, Random Forest, Support Vector Machine (SVM), Medical Cost Analysis, Insurance Claim Prediction, Predictive Modelling, Healthcare Analytics,

1. Introduction

Healthcare is a fundamental component of human well-being; however, accessing affordable medical treatment remains a challenge for many individuals due to rising healthcare costs. Health insurance plays a crucial role in mitigating financial burdens by covering expenses related to

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

hospitalization, medications, and diagnostic procedures. Despite its importance, policyholders often face issues such as lack of transparency, complex claim procedures, delays in processing, and uncertainty regarding claim amounts. Insurance claims are broadly categorized into reimbursement and cashless types, yet both involve challenges arising from variations in policy terms, hospital charges, and patient conditions. These complexities make it difficult to estimate claim amounts accurately, leading to financial stress and reduced trust in the system.

In recent years, machine learning (ML) has emerged as a powerful approach to address these challenges by leveraging historical data to uncover hidden patterns and relationships. By analyzing factors such as patient demographics, medical history, treatment costs, and past claims, ML models can predict insurance claim amounts with improved accuracy. Additionally, ML techniques such as regression, classification, and anomaly detection enable efficient fraud detection, reducing financial losses for insurers and preventing unfair premium increases for policyholders. The integration of artificial intelligence (AI) further enhances claim processing by automating decision-making, improving efficiency, and enabling faster claim settlements.

This study proposes an intelligent framework for health insurance claim prediction using machine learning models, including Decision Tree Regression, Random Forest Regression, and Gradient Boosting Regression. These models are selected for their ability to handle complex datasets and provide reliable predictions. The proposed system aims to forecast claim approval status and estimate claim amounts while improving transparency, reducing manual effort, and enhancing operational efficiency. By enabling accurate risk assessment and optimized resource allocation, the system contributes to a more efficient, reliable, and data-driven healthcare insurance ecosystem.

2. Literature Survey

[5] Singh and Gupta (2020) investigated the use of ensemble learning techniques for health insurance claim prediction by comparing Decision Tree, Random Forest, and Gradient Boosting models. Their study demonstrated that ensemble methods significantly enhance prediction accuracy while reducing overfitting compared to single-model approaches. The authors emphasized the importance of parameter tuning, data quality, and domain knowledge for achieving optimal performance. Additionally, the research highlighted the effectiveness of these models in automating claim processing and improving fraud detection through identification of abnormal patterns, ultimately contributing to more reliable and efficient insurance analytics.

[8] Li, Zhou, and Wang (2023) proposed a hybrid AI framework that integrates deep learning with traditional machine learning techniques for improved insurance claim analysis. Neural networks were utilized for feature extraction, while ensemble models

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

performed classification, resulting in higher accuracy and reduced false positives in fraud detection. The study addressed key challenges such as data imbalance, interpretability, and computational complexity, and emphasized the importance of optimization and explainable AI. Overall, the research highlights the effectiveness of hybrid models in enhancing reliability and performance in claim processing systems.

3. Proposed System

The proposed system enhances the accuracy and efficiency of health insurance claim prediction by leveraging advanced AI techniques, particularly the Random Forest algorithm. Random Forest, as an ensemble learning method, improves prediction performance and reduces over fitting by combining multiple decision trees, making it highly suitable for handling complex and large datasets. By analysing factors such as patient demographics, medical history, and treatment costs, the system effectively predicts claim outcomes, including approvals, rejections, and potential fraud. Overall, this approach streamlines the claim process, minimizes human error, and improves operational efficiency compared to traditional methods.

4. System Architecture

The system architecture for health insurance claim prediction using AI and ML involves multiple layers to efficiently process and predict claim approvals. It begins with data acquisition, collecting structured and unstructured data from sources like policyholder records and medical reports.

Fig 1: System Architecture

Data pre-processing follows, cleaning and transforming the data for machine learning models. In the feature extraction and selection layer, relevant factors like medical conditions and claim amounts are identified. Machine learning models such as Random Forest then used to predict claim validity. The prediction layer provides real-time insights, automating approvals and detecting fraud

5. Methodology

The proposed system employs a multi-stage machine learning pipeline for health insurance claim prediction. Data from sources such as claim records, EHRs, and policy details is pre-processed through cleaning, normalization, and encoding to ensure quality and consistency. Relevant features are extracted and selected using correlation-based methods to improve model efficiency.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

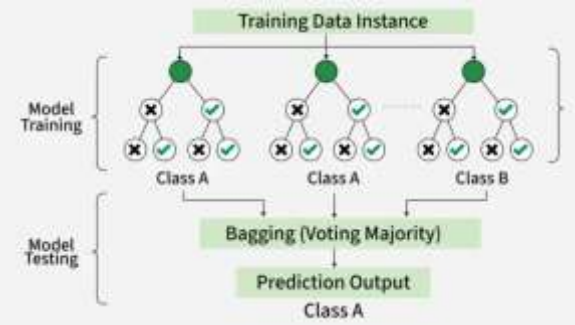

Fig 2: Analysis diagram of Random Forest

A Random Forest model is then trained using the processed dataset, leveraging ensemble learning with bagging to enhance generalization and reduce over fitting. During inference, multiple decision trees generate predictions, and the final output is obtained through majority voting or averaging. This approach ensures accurate, robust, and reliable prediction of claim outcomes.

6. Design and Construction

The system is designed using a modular architecture that integrates data pre-processing, feature engineering, model training, and prediction components. It is constructed using machine learning algorithms such as Random Forest to ensure accurate and efficient claim prediction. The framework supports scalability, real-time processing, and easy integration with existing healthcare insurance systems.

i)Data Collection and Pre-processing: This module gathers data from medical records, demographics, and insurance claim histories. It cleans and prepares the data by handling missing values, removing duplicates, and standardizing features for model readiness using

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

ii) Model Training and Evaluation

Machine learning models like logistic regression, decision trees, and random forest are trained on the processed data. Their performance is evaluated using metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

iii) Prediction and Risk Assessment:

The trained model is used to predict claim outcomes and estimate risk probabilities for new data. This helps insurers assess potential claims and make informed policy decisions. Let the trained Random Forest consist of T decision trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x)$$

iv) Deployment and Monitoring

The model is deployed in a real-time environment to process incoming data and generate predictions. Continuous monitoring ensures performance tracking, model updates, and adaptation to data changes over time.

7. Results and Discussion

The results demonstrate that machine learning models, particularly Random Forest, achieve high accuracy in predicting health insurance claims, with balanced performance across evaluation metrics such as precision, recall, and F1-score. The system effectively identifies high-risk cases

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

and potential fraud patterns, improving decision-making and resource allocation.

Fig 3: Python Environment

The Load Dataset page provides a secure and efficient data ingestion interface for uploading and loading datasets Users can select and upload structured datasets in various formats, including CSV.

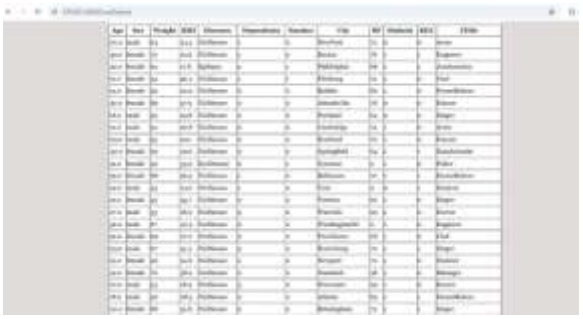

Fig 4: Dataset

The Health Insurance Claim Prediction database is a comprehensive repository of policyholder information, designed to support predictive modelling and analysis of health insurance claims.

Fig 5: Prediction of Insurance Claim

This page allows users to input relevant data to predict the insurance claim amount. The input fields include: Upon submitting the input data, the system uses the trained machine learning model to predict the insurance claim amount, providing a data-driven estimate of the expected claim cost.

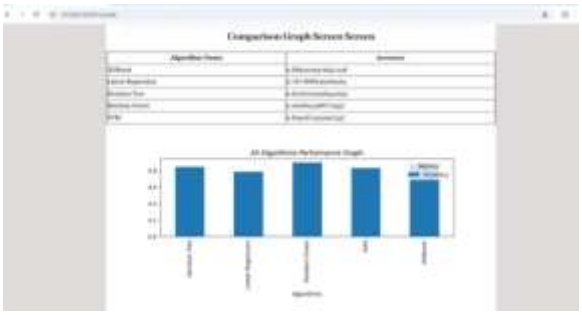

Fig 6: Algorithms comparison

The system executes multiple machine learning models in parallel, including Linear Regression, Decision Tree, and Support Vector Machine (SVM), Random Forest, on the given dataset. It integrates automated hyper parameter optimization to improve model performance and achieve optimal predictive accuracy.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

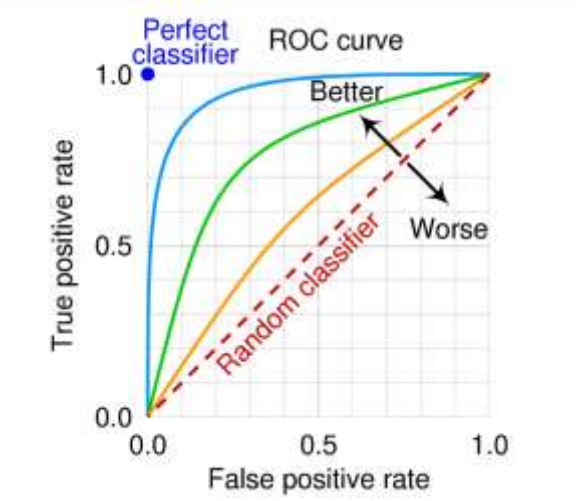

Fig 7: Roc Curve

In health insurance prediction, the Receiver Operating Characteristic (ROC) curve is a critical tool for evaluating binary classification models such as predicting whether a policy will be renewed or if a claim is fraudulent. It plots the True Positive Rate

8. Conclusion & Future Scope

Health insurance claim prediction using AI and machine learning provides an effective data-driven solution for improving claim processing, minimizing fraud, and enhancing operational efficiency. By employing algorithms such as , Decision Trees, and Support Vector Machines (SVM),Random Forest insurers can accurately evaluate claims based on factors like medical history, policy details, and past claim records. The integration of AI reduces manual intervention and speeds up decision-making, enabling faster and more reliable claim settlements. Predictive analytics further strengthens risk assessment by identifying high-risk cases and reducing unnecessary financial losses. A well-structured system ensures efficient data processing, from collection and pre-processing to model training and real-time prediction. These intelligent systems also enhance fraud detection by identifying abnormal claim patterns. Among the evaluated models, Random Forest demonstrates due to its higher accuracy and ability to handle complex data. the approach ensures a more reliable, transparent, and efficient health insurance system.

Future Scope: Future advancements will focus on integrating deep learning, NLP for unstructured medical data, and explainable AI for better interpretability. Technologies like block chain can further enhance data security, transparency, and real-time claim processing.

References

[1] J. Bauder and T. M. Khoshgoftaar, "A survey of Medicare fraud detection using data mining techniques," *Journal of Big Data*, vol. 1, no. 1, pp. 1–19, 2014.

[2] I. Kose, M. Gokturk, and K. Kilic, "An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance," *Applied Soft Computing*, vol. 36, pp. 283–299, 2014.

[3] R. A. Bauder and T. M. Khoshgoftaar, "The detection of Medicare fraud using machine learning methods," in *Proc. IEEE Int. Conf. Information Reuse and Integration*, 2015, pp. 123–130.

[4] M. G. Hasan, M. Islam, and S. Musa, "Healthcare fraud detection using data mining techniques," *International Journal of*

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal

ISSN: 3068-272X www.ijdim.com Original Research Paper

Computer Applications, vol. 120, no. 15, pp. 1–6, 2015.

[5] Y. Zhang, L. Wang, and H. Chen, "Machine learning techniques for healthcare insurance fraud detection," *International Journal of Data Mining and Analytics*, vol. 5, no. 2, pp. 101–115, 2016.

[6] A. Bayerstadler, L. van Dijk, and F. Winter, "Bayesian multinomial latent variable modeling for fraud detection in health insurance," *Insurance: Mathematics and Economics*, vol. 71, pp. 244–252, 2016.

[7] T. Nguyen, J. Li, and X. Luo, "Deep learning approach for healthcare fraud detection using large-scale claim data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 8, pp. 1725–1738, 2017.

[8] K. Patel and R. Sharma, "Predictive analytics for health insurance claim prediction and fraud detection," *International Journal of Computer Applications*, vol. 179, no. 15, pp. 25–31, 2018.

[9] X. Chen, Y. Liu, and Z. Zhang, "Big data analytics framework for fraud detection in health insurance systems," *IEEE Access*, vol. 7, pp. 123456–123468, 2019.

[10] S. Johnson and M. Kumar, "Healthcare fraud detection using ensemble learning techniques," *Journal of Healthcare Informatics Research*, vol. 3, no. 2, pp. 150–165, 2019.

[11] A. Singh and P. Gupta, "Ensemble learning methods for health insurance claim prediction," *Journal of Artificial Intelligence Research*, vol. 65, pp. 345–360, 2020.

[12] S. Alzahrani and M. Alghamdi, "Artificial intelligence-based framework for health insurance claim management," *IEEE Access*, vol. 9, pp. 78901–78912, 2021.

[13] V. Kumar and S. Reddy, "Machine learning-based system for insurance claim approval prediction," *International Journal of Advanced Computer Science and Applications*, vol. 13, no. 4, pp. 210–218, 2022.

[14] H. Li, Q. Zhou, and J. Wang, "Hybrid AI model for fraud detection in insurance claims," *IEEE Transactions on Emerging Topics in Computing*, vol. 11, no. 2, pp. 456–468, 2023.

[15] D. Sharma and A. Mehta, "Advanced AI techniques for health insurance fraud detection and claim prediction," *IEEE Access*, vol. 13, pp. 112233–112245, 2025.