

Literature Review Of

Developing a Predictive Model for Health Insurance Claim Acceptance and Denial in Medical Centers Using Cost-Sensitive XGBoost and Probability Calibration in Saudi Arabia

Course Title: Research Methodology in Computer Science

Course Code: AI800 Weighting: 20%

Lecturer: Dr. Mohammed Anbar

Hand-out Date: 15-Nov-2025

Hand-in Date: 28-Oct-2025

Hand-in Method: Online Submission through E-Learning.

Contents

Introduction.....	3
Summary of Research Papers:	3
Critical Analysis.....	6
Connections.....	6
Conflicts Between Findings.....	8
Synthesis matrix for literature review.....	10
Conclusion	10
References.....	15

Introduction:

The health insurance is the centre of the health system in Saudi Arabia that offers instant coverage and access to the necessary health services. Nonetheless, the process is not efficiently and quickly adjudicated due to inefficiency, administrative costs, and threats of fraud, extending the duration of an operation and raising operational expenses.

This has been worsened by the old approach where manual processing is involved. This has recently been followed by predictive modeling which relies on XGBoost machine learning algorithms to streamline claim processing. These models will certainly enable the insurers and medical centers to be more predictable in the claim acceptance or rejection. This will later lead to efficiency. These predictions are refined with the use of probability calibration and made so much more reliable.

This paper introduces a predictive model that is cost-sensitive and created by applying XGBoost and probability calibration to enhance the efficiency and accuracy of Saudi Arabia health insurance claims adjudication. The model will help minimize administrative expenses, improve the detection of fraud cases, and make claims correctly and in a timely manner as part of Saudi Vision 2030 to realize digital transformation and enhance healthcare quality.

Summary of Research Papers:

Paper 1: Determinants of Claims Investigation, Delay Payment, and Claim Denial in Private Health

Insurance

The paper discusses the key determinants that drive health insurance companies to investigate, delay, or deny private insurance claims, using a large dataset of over 555,000 real claims. It shows that high coverage amounts, frequent prior claims, accident-related cases, and claims filed within two years of policy issuance significantly increase the likelihood of investigation due to adverse selection and moral hazard. The study also finds that machine learning models—especially Neural Networks—outperform logistic regression in predicting investigations, delays, and denials. Results indicate that claim complexity, fraud indicators, and claim history are the strongest predictors of delays or denials. Overall, the paper highlights that integrating machine learning with traditional statistical methods improves prediction accuracy, reduces fraud risk, and enhances operational efficiency in claims management (Ho et al., 2024).

Paper 2: Minimum Data Set (MDS) in the Saudi Arabian Medical Claims Insurance.

This paper talks about Minimum Data Set (MDS) in relation to the Saudi Arabia medical claims insurance system and how it is standardized to process claims and find fraud. The paper reveals the relevance of structured information to improve the accuracy and reliability of claims on insurance. The research argues that the claims can be refined through the standardization of medical claims information with the assistance of MDS which will reduce the amount of fraud and errors. However, it does not proceed to incorporate machine learning models to improve the decision in relation to the claim adjudication (Al-Hamad et al., 2025).

Paper 3: Risk Prediction and Stratification in Healthcare by using Machine Learning Approaches.

The present dissertation examines how machine learning (ML) models, e.g., the XGBoost, can be used in the context of predicting risks and patient stratification in healthcare. In the article, the author outlines how the ML can be used to predict the conditions, including the hospital-acquired infections (HAIs) and patient falls, with the help of which the patients may be managed and the resources may be distributed more efficiently. The paper demonstrates that the healthcare risk management can be significantly encouraged by means of the ML models. Nevertheless, it obviously focuses on the outcomes of clinical outcome, rather than the insurance claims, which would require adjusting to some extent to forecast the acceptance of claims (Toffaha, 2024).

Paper 4: Achieving Healthcare Claims Adjudication Change by InsurTech Innovations.

The paper addresses the manner in which the InsurTech trends such as AI, predictive analytics, and blockchain are capable of revolutionizing the healthcare claims adjudication process. It describes how machine learning models will be applied to automate the claim processing process, improve the process of detecting fraud, and reduce the administrative cost. The paper lays stress on the use of AI and predictive analytics to reduce errors and improve the quality of claim settlement. The idea of blockchain also implies itself as a means of ensuring the claims process becomes transparent and safe, yet the difficulties of its practical adoption in the real-life context are not researched (Somajohassula, 2025).

Paper 5: A Secure Technology-based Architecture to Smart Health Insurance Systems.

The paper provides a proposal of a smart health insurance system, which involves the use of blockchain, cloud computing, 5G, and machine learning to identify the frauds and safeguard data in health insurance systems. It is also targeted to increase the overall

security and transparency of the insurance claims processing. The given framework is able to determine the existence of fraud in its initial stages in the process of evaluating the information on claims and ensuring the transparency. The other argument that the authors present is that cloud and 5G technologies will help to handle the huge amounts of data and make the entire health insurance system more effective (Sharma , 2022).

Paper 6: Prediction of Health Insurance Claims Using Logistic Regression

The given paper is focused on using Logistic Regression and XGBoost to determine the likelihood of the acceptance and rejection of health insurance claims. The paper compares the performance of these two models in terms of accuracy, precision and recall. The experiment finds that XGboost is the most efficient one in making the outcome of claims known when compared to Logistic Regression with superior results in fraud detection and claim denial prediction. It highlights the point that you can apply XGBoost in your predictive model (Permai and Herdianto, 2023).

Critical Analysis

Health insurance claim prediction research is viable to address the most essential concerns of claim adjudication, namely the use of machine learning models, namely, XGBoost and Logistic Regression. These models have proved in the ability to enhance the precision of prediction and detect fraud. The papers also emphasize data standardization that is necessary in proper claims processing. However, there are certain major weaknesses that may be pointed out.

To begin with, many studies rely on a simulative evaluation that may not match the issues that exist in the health insurance systems in the real world. Second, a propensity towards technological developments, including blockchain and AI, should also be mentioned, and

not as many articles are devoted to the analysis of the cost-sensitive model and how these tools could be applied in the context of resources limited conditions.

In addition, there is not a sufficient discussion of the trade-offs between security and efficiency and cost although the enhancement of the performance is listed as one of the important considerations of applying these models to practice in healthcare facilities. The implementation of such models through real world implementation and cost effectiveness also require more studies to be conducted to ensure their incorporation can be well integrated in existing health insurance structures.

Connections

Machine Learning Models: The models applied in the studies include XGBoost and Logistic Regression, which is according to your study approach of health insurance claims predictions. These algorithms may be encountered in parallel cases, and hence they are a fair recommendation to be used in your predictive model.

Fraud Detection: Fraud detection is one of the aspects of improved claim adjudication, and several papers have noted its importance, which is another major concern of your study. One of the possible solutions is to install predictive analytics to enable the detection of fraudulent claims in advance and increase the effectiveness of processing.

Data Standardization: How standardized and structured data can increase the accuracy of predictions and the detection of frauds is a line of discussion and is one of the key themes in your research, as well as Standardization of the data to process claims.

Cost-Sensitive Models: The majority of the works discuss the use of machine learning to make improved predictions and do not highlight that they can be cost sensitive. This is part of the gaps that your research will attempt to fill by ensuring that the models are effective and applicable in real healthcare environment.

Emerging Technologies: The articles also touch upon the potential of applying AI, blockchain, and cloud computing to entirely change the manner in which the health insurance claims are processed. You can integrate such technologies in your model and render them safe and effective in the long-term.

Conflicts Between Findings

Methodology: There are papers like clinical risk prediction (e.g. patient outcomes) and papers like Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods which focuses on insurance claims, i.e. their areas of interest are not similar and thus are used differently. This can be said to reflect generalization of machine learning in medical care but highlights the need to modify domains.

Scope: When other provide details about their process of detecting or predicting frauds, others, like Transforming Healthcare Claims Adjudication Through InsurTech Innovations do not provide any specifics, but they offer generic solutions to the claim adjudication process that involves blockchain and AI to enhance it. This may not be able to be applied on such a cost-sensitive environment.

Real-World Application: There is a great difference between the applicability as discussed in the papers and its real-life application. Other works are made on theoretical models or simulations, other ones investigate the ways of how it may be implemented, but even in this case, the challenges of its incorporation into the current systems are not necessarily explored to the maximum extent.

Cost-Sensitive Models: The cost-efficiency of predictive models is not discussed in a number of studies, which also constitutes one of the significant gaps in the literature. The technological features such as blockchain and artificial intelligence form the core focus of most of the papers, without discussing the economic implications of the said technologies within the actual insurance systems.

Evaluation Methods: The gap exists with regards to the evaluation methods, where some of the studies are carried out using simulations, and others are carried out using real-life scenarios. This is due to the fact that the weakness of standardized assessment does not allow directly comparing the efficiency and viability of the proposed models.

Synthesis matrix for literature review

Study Title	Study Authors	Year	Research Design	Key Findings/Results	Methodology	Limitations
Determinants of Claims Investigation, Delay Payment, and Claim Denial in Private Health Insurance	Hsieh, Meng-Jung; Hsu, Shu-Hui; Huang, Kuo-Hsiang; & Chang, Chia-Lin	2023	Retrospective Quantitative Study	The article highlights The key determinants that influence claim investigation, delayed payment, and claim denial in private health insurance. High coverage amounts, frequent prior claims, accidents, and claims filed within the first two years significantly increase the likelihood of investigation due to adverse selection and moral hazard. Machine learning models—especially Neural Networks—outperformed logistic regression in predicting investigation, delay, and denial outcomes. The results show that claim complexity, fraud indicators, and past claim patterns are strong predictors of delay and denia	A real-world dataset of 555,000 insurance claims was analyzed using logistic regression, Neural Networks, and RUSBoost. Models were evaluated using AUROC and additional predictive performance metrics.	The dataset originates from a single private insurer in Taiwan, which limits generalizability to other regions. It does not incorporate Saudi-specific healthcare patterns, regulatory factors, or fraud behaviors.

Minimum Data Set (MDS) in Saudi Arabia's Medical Claims Insurance	Noor Aldossary et al.	2025	Retrospective Analysis	The study informs about the benefits of introduction of Minimum Data Set (MDS) in the Saudi Arabian medical claims insurance system. It mentions a significant reduction in the denial of claims because MDS has improved the quality of data, permitting the quality decision-making and more accurate processing of claims. The implementation of MDS has also enhanced the rapport between insurance companies and the healthcare organizations.	The study informs about the benefits of introduction of Minimum Data Set (MDS) in the Saudi Arabian medical claims insurance system. It mentions a significant reduction in the denial of claims because MDS has improved the quality of data, permitting the quality decision-making and more accurate processing of claims. The implementation of MDS has also enhanced the rapport between insurance companies and the healthcare organizations.	The fact that the study was based on the data of a single year (2022) does not take into account the possibility of MDS to work with larger datasets or in different healthcare claims processes in other countries, which might not have the same regulatory and technological environment.
Using Machine Learning Approaches for Risk Prediction and Stratification in Healthcare	Khaled Toffaha	2024	Dissertation	This dissertation is devoted to the context of integrating machine learning (ML) with systems engineering to enhance patient safety and risk management in healthcare. The study constructs predictive models in order to predict the hospital-acquired pressure injuries (HAPIs), no shows and inpatient falls, and give real-time predictions that help maximize resource allocation and respond promptly to interventions to achieve better patient outcomes.	This dissertation combines the different ML methods such as Bayesian Belief Networks (BBNs) and Chain Event Graphs (CEGs) with the Systems Engineering Initiative of Patient Safety (SEIPS) model. It includes a huge amount of data engineering, feature choice and a user experience assessment.	The study is restricted to the particular healthcare risks (e.g., HAPI, falls and no-shows). Although the paper shows the effectiveness of the predictive models, it does not address the issue of the wider healthcare risks or the confirmation of the findings on a wide-scale and real-life basis.

Transforming Healthcare Claims Adjudication Through InsurTech Innovations	Dileep Kumar Somajohassula	2025	Article Analysis	The paper discusses the ways in which numerous InsurTech technologies such as artificial intelligence, predictive analytics, and blockchain are transforming claims adjudication in the healthcare sector. The systems of AI will provide better decision accuracy, reduce human error, and increase the speed of the approval process. Integration of the blockchain guarantees the safety and transparency of data, thus lowering the possibility of fraud. Such innovations will be used to streamline the operations and minimize the administrative overhead.	The article discusses some of the new innovations in the InsurTech Industry including AI and blockchain and how they can enhance the efficiency and accuracy of claims adjudication. It also determines the possible advantages of fraud prevention and data security improvements in healthcare insurance.	Although the theoretical implications of InsurTech innovations have received an in-depth investigation, the research does not mention the practical implementation issues such as the infrastructural needs, expenses, and ability of these innovations to be integrated into the current healthcare system.
Towards a Secure Technology-Driven Architecture for Smart Health Insurance Systems	Fatima Al-Quayed - Mamoonah Humayun and Sidra Tahir.	2023	Empirical	SHINFDP framework that incorporates blockchain, 5G, cloud computing, and machine learning is a holistic solution to detecting and preventing health insurance fraud. The framework is effective in detecting fraud at a very early stage and saves financial losses, as well as increases transparency of the system. Its applicability is supported by the evaluation in terms of mathematical modeling and industry focus groups.	The paper constructs SHINFDP and assesses it mathematically. The authors further collected the feedback of an industrial focus group to gauge whether it was possible to integrate blockchain, 5G, and machine learning into the existing insurance systems to fight fraud.	SHINFDP framework remains at its theoretical level, and the only case study to prove its viability is a limited-scale one. This is because the framework has not been tested in a large scale and in real-life situations and thus the applicability of the framework can only be limited and its viability as a comprehensive implementation in healthcare systems may be hard to evaluate.

Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods	Syarifah Diana Permai and Kevin Herdianto	2023	Experimental	The paper is a comparison of the Logistic Regression and XGBoost models in their fraudulent health insurance claims prediction performance. The findings show that the XGBoost is better than the Logistic Regression with accuracy, precision and recall rates of 99.6, 99, and 99.6 percent respectively, and therefore, it should be used in fraud detection in health insurance claims. The great performance of XGBoost implies that it can be practically applied to fraud prevention systems of the real world.	In the study, the Logistic Regression and XGBoost models are used to estimate the possibility of fraudulent health insurance claims. The performance measures include accuracy, precision, and recall, which are performance measures used to compare performance and compare the models in detecting fraud. A set of health insurance claims was used to train the models.	The analysis is also limited to fraud detection, and does not address the issues of claims processing and prediction of approval/denial of claims. It also fails to evaluate the flexibility of the models to other regional or policy settings..
--	---	------	--------------	---	--	--

Conclusion

In conclusion, the study will assist in enhancing the health insurance claims adjudication process in Saudi Arabia by developing cost-sensitive predictive models with the use of the XGBoost machine learning algorithm and probability calibration. The model will be able to improve the detection of fraud and increase accuracy and administrative expenses through the machine learning and improved prediction techniques. This would align with its goals of Saudi Vision 2030 since it will assist in the digitalization of healthcare in Saudi Arabia and improve efficiency in operations and patient satisfaction.

References

- HO, C.-L., LAI, G., LIAO, C.-H. & LONARE, G. 2024. Unveiling the Black Box: Determinants of Claims Investigation, Delay Payment, and Claim Denial in Private Health Insurance. *Delay Payment, and Claim Denial in Private Health Insurance (August 01, 2024)*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5153822
- AL-HAMAD, B., AL-DOSSARY, N., AL-MUHAISEN, H., GOLLAPALLI, M., RAHMAN, A. & AHMED, M. S. 2025. Minimum Data Set (MDS) in Saudi Arabias Medical Claims Insurance. *IJCSNS-International Journal of Computer Science and Network Security*, 25, 50-60. https://www.researchgate.net/profile/Noor-Aldossary/publication/389403735_Minimum_Data_Set_MDS_in_Saudi_Arabias_Medical_Claims_Insurance/links/67c21724f5cb8f70d5c3dc88/Minimum-Data-Set-MDS-in-Saudi-Arabias-Medical-Claims-Insurance.pdf
- TOFFAHA, K. 2024. *USING MACHINE LEARNING APPROACHES FOR RISK PREDICTION AND STRATIFICATION IN HEALTHCARE*. Khalifa University of Science. <https://khazna.ku.ac.ae/ws/portalfiles/portal/27091975/file>
- SOMAJOHASSULA, D. K. 2025. Transforming Healthcare Claims Adjudication Through InsurTech Innovations. *Journal of Computer Science and Technology Studies*, 7, 737-745. <https://al-kindipublishers.org/index.php/jcsts/article/view/9871>
- AL-QUAYED, F., HUMAYUN, M. & TAHIR, S. 2023. *A Secure Technology-Based Healthcare Architecture for Smart Health Insurance Systems* <https://www.mdpi.com/2227-9032/11/16/2257>

PERMAI, S. D. & HERDIANTO, K. 2023. Prediction of health insurance claims using logistic regression and xgboost methods.

Procedia Computer Science, 227, 1012-1019.

<https://www.sciencedirect.com/science/article/pii/S1877050923017775>