Computer Science and Information Technologies
Vol. 6, No. 3, November 2025, pp. 346~354
ISSN: 2722-3221, DOI: 10.11591/csit.v6i3.p346-354

Predictive model for high-risk healthcare clients and claims frequency

Lenias Zhou, Mainford Mutandavari, Lucia Matondora
School of Science Information and Technology, Harare Institute of Technology, Harare, Zimbabwe

Article Info

Article history:

Received 7 Jun, 2025
Revised 27 Jun, 2025
Accepted Jul 3, 2025

Keywords:

Bayesian optimization algorithm
Claims frequency
Deep learning
Healthcare insurance
High-risk clients
Predictive modeling

ABSTRACT

Global healthcare spending surged to approximately USD 9.8 trillion in the aftermath of the COVID-19 pandemic, intensifying the need for effective risk management strategies in healthcare insurance. This study proposes a predictive model designed to identify high-risk clients for timely targeted interventions and to forecast claims frequency for optimized resource allocation. A real-world claims dataset from a healthcare insurance provider was utilized. Bayesian optimization was employed to enhance data labelling. A deep learning (DL) model with sigmoid activation was used to classify high-risk clients, while a regression model forecasted claims frequency. The model was trained and validated, and gave an accuracy of 97%, a precision of 95.2%, a recall of 98.1% and an F1-score of 96.6%. The results confirmed the model's accuracy in identifying high-risk clients and its ability to provide reliable forecasting of future claims frequency. Importantly, the model also provided the reason behind its classification decision, enhancing transparency and trust. This research provides valuable data-driven insights to both the healthcare insurers and clients, giving them the power to stay ahead in managing key risks, which ultimately reduces the cost of healthcare insurance. This work contributed a scalable and interpretable solution for risk prediction in healthcare insurance.

Corresponding Author:

Lenias Zhou
School of Science and Information Technology, Harare Institute of Technology
Belvedere, Harare, Zimbabwe
Email: zhoulenias@gmail.com

1. INTRODUCTION

In today's dynamic business environment, sustainability and profitability are closely tied to effective risk management, particularly within the healthcare insurance sector. The early identification of high-risk clients is no longer a luxury but a necessity, enabling insurers to mitigate financial strain and improve health outcomes, allocate resources strategically, and set premiums accurately. However, traditional risk identification methods often fall short in addressing the complexities of modern healthcare insurance landscapes. They typically focus on predicting high-cost clients without offering insights into claims frequency, which is equally critical for comprehensive risk profiling and financial planning. Furthermore, they are frequently challenged by overfitting, limited interpretability, and reliance on small or synthetic datasets, which undermine their reliability. This paper proposed a predictive model that integrates advanced machine learning (ML) techniques to enhance both the precision of risk identification and the forecasting of claims frequency. By leveraging robust data-driven methodologies, the model aims to support more informed decision-making and foster resilience in healthcare insurance operations.

2. RELATED WORK

There are works that were done by other researchers on the application of predictive analytics within the domains of business and healthcare. These prior studies employed diverse methodologies and encountered a range of challenges. Table 1 presents a synthesized overview of selected studies along with their respective limitations.

Table 1. Summary of prior works on predictive analytics in health insurance

Focus area	Reference authors	Common merits	Common limitations
Predictive analytics in business and healthcare	[1]–[3]	Broad application. Improved outcomes. Strategic insights.	Lack of empirical validation evidence. High deployment costs. Data quality issues
Health insurance costs and claim prediction	[3]–[7]	High prediction accuracy. Wide model comparison. Ethical AI integration.	Overfitting, limited interpretability, and the use of small and synthetic datasets.
Risk assessment and high utilization prediction	[6], [8]–[12]	Real-world relevance, large datasets, and improved risk prediction.	Resource-intensive, missing variables, and limited model diversity. Few risk factors were considered due to system limitations. Noise from oversampling.
Systemic reviews and comparative studies	[13]–[16]	Comprehensive model coverage, time series forecasting, and real-world data usage.	Limited explainability. Used short timeframes and high sensitivity to outliers. High risk of bias. Limited clinical implementation.
Advanced AI integration in healthcare	[17], [18]	High accuracy and real-time updates. Multimodal data integration.	Regional data limitations and imbalances in rare conditions.
Fraud detection in healthcare	[19]	High fraud detection accuracy on the use of models.	Small sample size. Limited number of features. Manual feature engineering. Handling of gender imbalance in datasets.
Time series statistical analysis of claims	[20]–[22]	Better handling of skewed data.	Challenges with censored data. Could not be generalized. Data quality issues.
Responsible AI in healthcare	[6], [23]	Comprehensive models evaluation on ethics	Used small samples with limited variables. Generalized ethical discussion. Scalability and equity concerns.

The research by Nwoke [1] and Nnamdi [2] examined the application of predictive analytics in decision-making to improve healthcare outcomes. Their findings can be generalized to most scenarios because there was a significant improvement in outcomes due to early detection and intervention [1], [2]. However, the approach lacked "empirical validation" due to high implementation costs [1], [2].

The work of Thakre *et al.* [3] centered on the prediction of insurance costs and fraud detection by employing ML models. Despite the high prediction accuracy that they attained, the models suffered from overfitting as well as a lack of interpretability [3], [20]. It is key to be able to know how the models are arriving at their decisions. Another, but lesser, criticism of their approach was the reliance on small and synthetic datasets, which Alam and Prybutok [6] identified as a major factor for overfitting. The application to real-world problems was illustrated in Ruijter *et al*. [9] and Li *et al*. [10] as they applied sizable datasets and achieved enhanced risk prediction. Both utilized ML models for risk stratification to pinpoint high-need patients.

The authors faced challenges in dealing with missing variables in the real-world medical data. The real-world medical data records demonstrate imbalances because high-cost scenarios are less frequent than standard and low-cost situations [9], [10], [12]. The other challenge that is common among these research papers is the issue of high computing power that was required to run the models. The study by Alotaibi [11] focused on the use of predictive analytics to identify risk factors within organizations. This work made remarkable progress in picking legal and regulatory risks as well as information technology risks, but there was a need for larger datasets. Alotaibi [11] used decision trees (DT), linear transformation (LT), and neural networks (NN) and pointed out that the approach was worth trying using other models.

Forecasting future health claims values was done in the work of Mashasha *et al* [13] using an autoregressive integrated moving average (ARIMA) model. They managed to forecast healthcare trends and future claim values from time series data. They had challenges with outliers. The forecasting approach used in [13] is ideally applicable to linear data and may not capture nonlinear, complex patterns.

Model comparison was also done in the work of [9] and [14]. The work of Aloyuni [14] first compared ML and deep learning (DL) and highlighted that ML required more feature engineering and assistance from the domain experts. On the other hand, DL could learn from raw data and was found to be more effective on large datasets [14]. In comparing convolutional neural network (CNN), support vector machine (SVM), generative adversarial network (GAN), and random forest (RF) across twenty publications, Aloyuni [14] found that CNN-based models, especially when combined with ensemble methods or GANs,

were frequently used for image-based diagnosis with a high accuracy and sensitivity. Hybrid models (for example, CNN+MAFW, CNN+GAN) were found to give improved performance as they combined feature selection, optimization, and classification techniques. Traditional models like SVM and RF were used in combination with DL for classification tasks or for structured data, but with overfitting, explainability, and preprocessing challenges [3], [9], [12], [14].

On the responsible use of artificial intelligence (AI) in healthcare, Akter *et al.* [23] warns about bias that can be amplified by AI models. This study pointed out that bias may lead to unequal treatment or misdiagnosis, particularly in underrepresented populations. The need for the development of AI models that are fair and inclusive was emphasized [17]. The electronic health records (EHR) contain sensitive personal data that requires strict protection to comply with regulations like the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR) [6], [7], [18]. The study by [3] and [9] advocates for explainable AI systems that allow users to understand how decisions are made. The researchers both concluded by recommending continuous refinement and monitoring of AI systems to maintain trust and effectiveness. The work of other researchers on the same note highlighted the need for structured frameworks and checklists to guide responsible deployment of AI models or systems [17], [23].

3. METHOD

A systematic approach was used to develop and evaluate a predictive model for identifying high-risk medical fund clients and forecasting claims frequency. The process involved collecting data, preprocessing, and exploratory data analysis (EDA), which included labelling the data, training the model, validation, and deployment. The design was correlational, with a predictive purpose. While traditional correlational research primarily seeks to determine the extent of relationships between two or more variables using statistical data, the ultimate objective here is to leverage these identified relationships to forecast future outcomes. This did not involve describing existing connections but quantifying them to enable reliable predictions of high-risk clients and forecasting claims frequency. The idea was to come up with an accurate model that is interpretable and scalable for real-world applications.

3.1. Our approach

The approach used in this study followed a deductive reasoning framework. We began by examining collected data variables and considering the existing theories. The concern was not about which variables are related but rather how strongly and in what direction these relationships exist to enable the generation of reliable future predictions. Figure 1 shows the overview of the research workflow that was followed.

3.2. Dataset

We used a real-world dataset containing five years (2020 to 2024) of healthcare claims. This dataset included patient demographics, treatment histories, and claim outcomes. The dataset had a total of thirty-five (35) features and nine hundred and thirty-four thousand eight hundred (934,800) claims from one hundred and ten thousand and three (110,003) unique members. The value of claims over the period amounted to one hundred and twenty-four million five hundred thousand United States dollars (USD 124,500,000). Zimbabwe operates under a structured currency system with the Zimbabwe Gold (ZiG) being the primary currency.

The company that provided us with the dataset prefers to report its financials in USD. If a person pays for services in local currency (ZiG), the amount is converted to the USD equivalent using the daily foreign exchange rate of that day for reporting purposes. This is the reason why our dataset value is stated in USD.

The handling of healthcare claims data is important for developing a robust model. The presence of irregular time series, high dimensionality, as well as privacy concerns requires careful data management. For this reason, the approach prioritized thorough data cleaning and strict ethical handling as much as, if not more than, the modelling process itself. The success of the predictive model hinged on effectively managing and transforming the complex, noisy, and sensitive real-world data. This approach examined the fields and grouped them according to their relevance for predictive modelling, as shown in Table 2.

Table 2 provides a clear and organized overview of the data fields that were fed into the predictive model. This listing of fields and explaining their relevance enhances the transparency and reproducibility of the research, allowing other researchers to build from the performed work. It also demonstrated a strong understanding of how healthcare data translates into meaningful predictors for both high-risk client identification and claims frequency forecasting [24].

From the 35 features in that dataset, our approach used three (3) compound features, which are direct indicators of potentially high-risk clients and can assist in forecasting claims frequency. The three compound features are the total amount claimed, the total amount rejected, and the number of claims per

month. The claims that financially strain the healthcare insurance providers are the large amounts and the frequency of the small, claimed amounts. The rejected claims also affect the client, who is a customer of the healthcare insurer and is of concern in flagging risk.

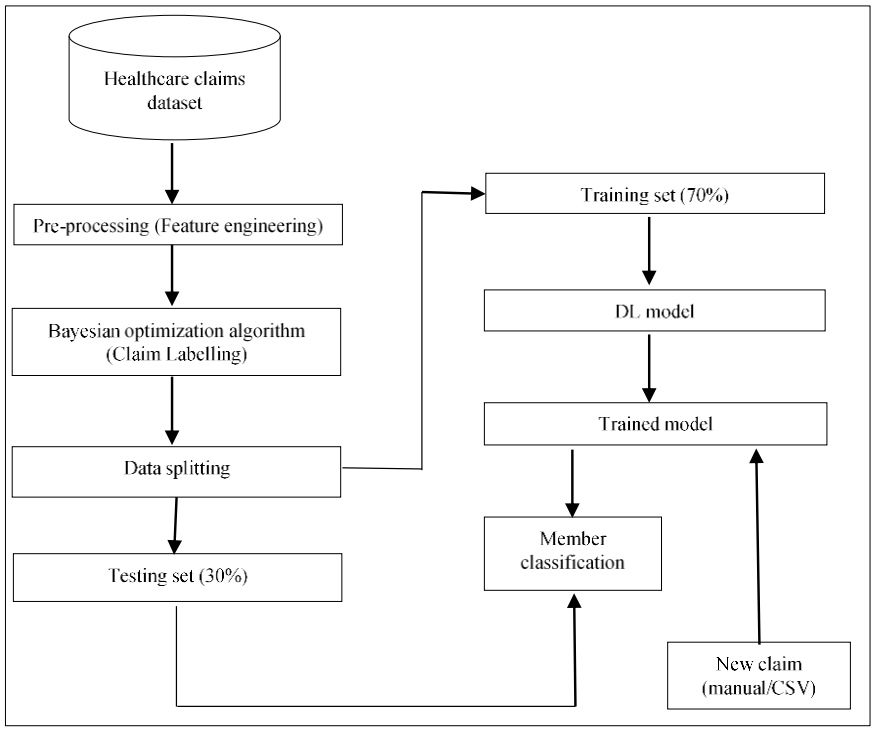

Figure 1. Research workflow

Table 2. Healthcare claims data fields and their relevance to predictive modelling

Category	Specific fields	Relevance for predictive modelling
Client demographics	Age, gender, race, address, deprivation index, and medical aid package.	Correlate with various health conditions and healthcare needs, influencing claims frequency, and risk stratification.
Claim details	Claim type (inpatient, outpatient, pharmacy, radiology, enrolment), provider identifier, and date of service.	Define the nature and timing of healthcare interactions, crucial for understanding service utilization patterns, and identifying specific claim types for analysis.
Clinical information	Diagnosis codes, procedure codes, total number of times diagnosed with target condition, lifestyle factors (smoking status, body mass index (BMI), and sporting activities).	Direct indicators of health status, disease burden, and types of service consumed, essential for risk assessment and predicting future claims. Comorbidity counts provide a measure of patient complexity [15].
Financial data	Billed amount, paid amount, rejected amount.	Directly inform the financial value of claims, crucial for forecasting costs, and identifying potentially high-cost clients.
Temporary data	Number of previous visits, time series of diagnostic history.	Provide longitudinal context for understanding patient behavior, disease progression, and predicting future utilization patterns [6].

3.3. Labelling process

We employed a Bayesian optimization algorithm to enhance data labelling, allowing efficient exploration of parameter spaces and enhancing the quality of labeled data for subsequent modeling. In this case, Bayesian optimization was being used beyond its traditional role; its principles (surrogate and acquisition function) were applied to select data points for labelling in active learning. An acquisition function checks for unlabeled data that will be most beneficial to label next for a ML model. Suppose you have a huge dataset but can only afford to have a small fraction of it labeled. The acquisition function guides in selecting the ones to label next and hence labels the whole dataset.

Bayesian optimization, in the context of "labelling ahead," is a "smart" curator of data where the high-risk healthcare clients are identified. It further directs the costly human labelling effort to the most valuable and informative data points, resulting in a more accurate and efficient DL model for risk classification at a significantly low labelling cost. The members within the dataset were labelled as high-risk if any of the following conditions were met:

– Total claimed amount>optimized threshold (typically around 75-90% percentile),
– Total rejected amount>optimized threshold,
– Claims Per Month>optimized threshold,

Otherwise, members were labelled low-risk.

Bayesian optimization was utilized to automatically find the best quantile thresholds. This approach improved reliability and reduced the need for subjective manual threshold settings. As a result, the model decisions are fully data-driven [25].

3.4. Classification process

A DL model was used, a simple, efficient feedforward NN. The network structure with five dense layers of "neurons" was used, starting with two hidden layers that process the input data using a common activation function called rectified linear unit (ReLU), which teaches the network to learn complex relationships [14]. The final layer, which makes the actual predictions, used a sigmoid activation function. The sigmoid activation function gives the overall result of either high risk or low-risk because it squashes its output into a number between 0 and 1, which can be directly interpreted as a probability [14].

The dataset of 934,800 medical claims instances was randomly separated into an evaluation dataset of 280,440 (30%) medical claim instances and 654,360 (70%) medical claim instances for training the model. The model used the Adam optimizer, which is like a smart teacher that adjusts how much the network learns from each mistake to make training faster and more efficient [14]. The training was done over 30 iterations (epochs), and after the training, the performance was checked using completely new, unseen data to give a true measure of real-world accuracy.

3.5. Model interface

To allow users to interact with the model and use it as a risk classifier tool, we created a user interface (UI) dashboard, as shown in Figure 2. The dashboard streamlines the process of dataset management, model training, and classification. The dashboard abstracts away the underlying code, making the powerful DL model accessible to users who may not have programming expertise, allowing them to leverage it for identifying high-risk clients and forecasting claims frequency. This makes the tool accessible to a broader audience.

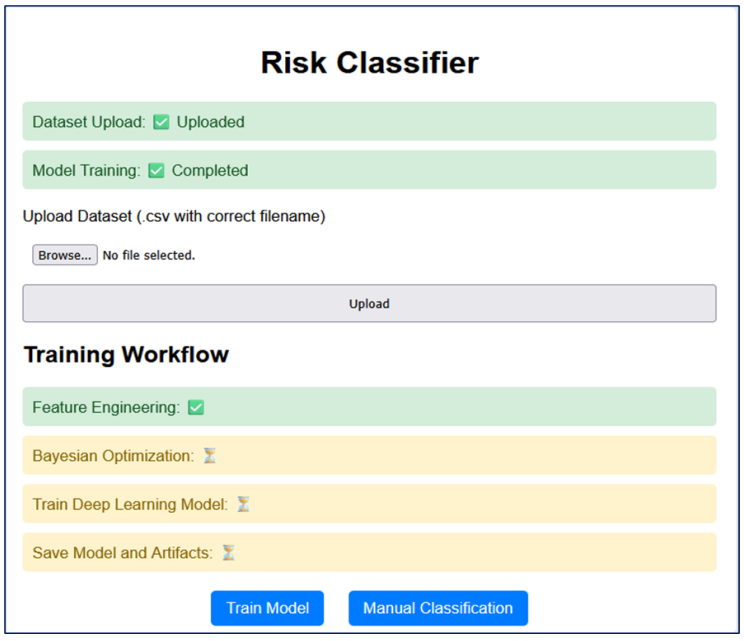

Figure 2. Risk classifier dashboard

The dashboard provides a user-friendly interface to monitor the status of the model; upload claim files in CSV format or manually enter the claim details using manual classification. It features pipeline status visualization for feature engineering, Bayesian optimization, DL model training, and confirms if the trained file has been saved successfully. It has action options for either retraining the model or manually classifying the claims.

4. RESULTS AND DISCUSSION

4.1. Training results

The DL model performed exceptionally well, correctly classifying 99.86% of unseen healthcare claims in the test dataset as either high-risk or low-risk. This showed a very strong predictive capability. The scaler was successfully saved, ensuring that the model can now be deployed and used without the need to be retrained. Both training and validation accuracy steadily increased and stabilized around 98-99.8%. The training and validation loss decreased significantly without overfitting, indicating that the model is well generalized. Figure 3 shows a screenshot of the model training results, where X-axis represents the epoch refers to the number of times the dataset has passed through the NN and Y-axis shows model accuracy as a percentage of correct predictions.

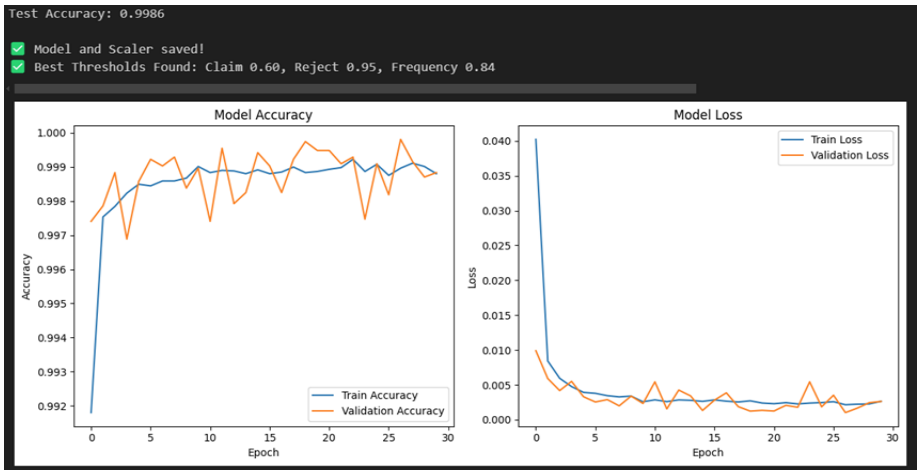

Figure 3. DL model training curves

4.2. Confusion matrix

The confusion matrix evaluates the model's ability to correctly classify high-risk and low-risk members as presented in Figure 4. The model predicted 18,602 clients as low-risk (0) when they were actually "low-risk" (0). This is an excellent performance improvement in identifying individuals who are genuinely not high risk. The model incorrectly predicted 43 clients as "high-risk" (1) when they were actually "low-risk" (0). These are type 1 errors. This may lead to unnecessary interventions for members who don't need such resource allocations. While 43 is a relatively low number compared to true negatives, the impact of false positives depends on the respective cost of such misclassification.

The model predicted 3 clients as "low-risk" while they were "high-risk" (1). These are type 11 errors. These are the most critical errors in this context. For healthcare, a false negative means that a high-risk individual goes unnoticed, potentially leading to adverse health outcomes, high future costs due to delayed interventions, or missed opportunities for preventative care. The fact that the number is very low is a strong positive for the model. The model also predicted 14,353 clients as "high-risk" when they were actually high-risk. This high number shows great performance for the model. This is the primary objective of this model is to identify high-risk clients.

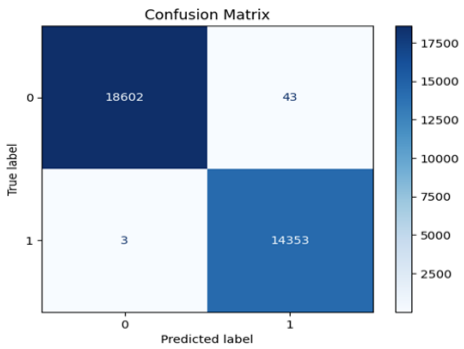

Figure 4. Model confusion matrix

4.2.1. High accuracy

Overall, most predictions fall into the true positive and true negative categories. This indicates a very low rate of misclassification. Such results are consistent with the high-test accuracy (0.9986%) reported earlier in section 4.1.

4.2.2. Minimizing critical errors

The most crucial aspect for identifying high-risk clients is minimizing false negatives, as missing a high-risk individual can have significant consequences. With only 3 false negatives, the model showed outstanding performance in this regard. This means very few genuinely high-risk clients are slipping through the cracks.

4.2.3. Manageable false positives

While there are 43 false positives, the number remains relatively small. Their impact should be considered carefully in the overall evaluation. This is especially true when weighed against the benefits of correctly identifying over 14,000 high-risk clients and nearly 19,000 low-risk clients.

4.3. Operational results

The following output is shown in Figure 5, demonstrates the real-world application of the DL model for healthcare risk identification and forecasting claims frequency. The model ingested new claim data seamlessly, with an option to upload a batch in CSV format. The model automatically processed data, giving a risk category to the member. The model then saved the result, making them available for further analysis or integration to other systems, and that's an "actionable output". The model also gave the claims frequency based on the available data, "predicted claims per month," offering a more direct quantitative forecast. The "explanation" part clarifies the rationale. This explanation for the decision enhances the practical utility of the model, allowing users to understand why a client has been flagged as high-risk, which is important for trust and effective intervention in healthcare. We used a manual, rule-based explanation approach due to its clarity and simplicity, as it explains the predictions in a way humans can understand. This empowers the users to act on these insights with greater confidence, as shown in the dashboard result in Figure 6.

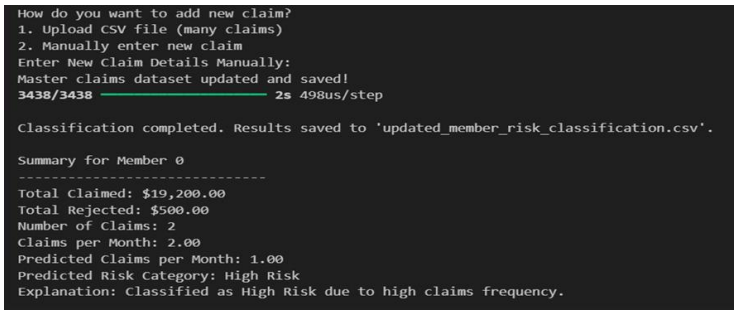

Figure 5. Model interaction outcomes in real-world applications

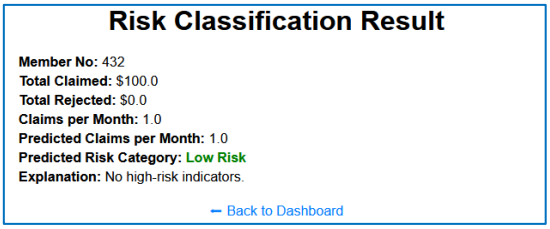

Figure 6. Dashboard classification result

5. CONCLUSION

The comprehensive work undertaken to come up with a predictive model for identifying high-risk healthcare clients and forecasting claims frequency has yielded exceptionally promising results, directly addressing all three core objectives. The primary objective of identifying high-risk clients has been achieved with outstanding accuracy, as evidenced by both testing and operational results. Additionally, it successfully forecasted claims frequency and provided an explanation for the predictions. These results are well supported

by the data and align with prior research, particularly in highlighting the three compound features (total amount claimed, total amount rejected, and claims frequency) as the critical indicators of high-risk behaviors in healthcare fraud detection. The model's precision surpassed typical expectations for such a complex predictive task, revealing its potential for proactive risk management in healthcare insurance. This shows potential to improve health outcomes and mitigate future healthcare costs. The other notable contribution is the integration of Bayesian optimization to enhance the performance of DL models on real-world healthcare claims data. The emphasis on explainability further presents this as a significant conceptual advancement in the application of DL to healthcare analytics. Future research could explore the integration of advanced costing mechanisms to support the deployment of a predictive model-as-a-service (PMaaS) platform. This will allow usage by individuals and small to medium-sized business enterprises who have initial investment capacity limitations.

FUNDING INFORMATION

The authors state that no funding was involved.

AUTHOR CONTRIBUTION STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Lenias Zhou	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓	
Mainford Mutandavari	✓	✓		✓	✓	✓		✓	✓	✓		✓	✓	
Lucia Matondora				✓	✓				✓	✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available on request from the corresponding author, [LZ]. The data, which contain information that could compromise the privacy of research participants, are not publicly available due to certain restrictions.

REFERENCES

[1] J. Nwoke, "Healthcare data analytics and predictive modelling: enhancing outcomes in resource allocation, disease prevalence and high-risk populations," *International Journal of Health Sciences*, vol. 7, no. 7, pp. 1–35, 2024, doi: 10.47941/ijhs.2245.
[2] M. Nnamdi, "Predictive analytics in healthcare," *ResearchGate*, Apr. 2024. [Online]. Available: https://www.researchgate.net/publication/379478196_Predictive_Analytics_in_Healthcare.
[3] V. P. Thakre, R. D. Poul, and A. D. Sawarkar, "Predictive precision: unraveling health insurance claim patterns with logistic regression and decision trees," *Cureus Journal of Computer Science*, 2025, doi: 10.7759/s44389-025-03010-y.
[4] A. A. Adesina, T. V. Iyelolu, and P. O. Paul, "Leveraging predictive analytics for strategic decision-making: enhancing business performance through data-driven insights," *World Journal of Advanced Research and Reviews*, vol. 22, no. 3, pp. 1927–1934, Jun. 2024, doi: 10.30574/wjarr.2024.22.3.1961.
[5] R. Soliman, "RWD107 better use of real-world data through predictive analytics: a knowledge-to-wisdom conceptual framework for evidence-based practice," *Value in Health*, vol. 25, no. 12, 2022, doi: 10.1016/j.jval.2022.09.2332.
[6] A. Alam and V. R. Prybutok, "Use of responsible artificial intelligence to predict health insurance claims in the USA using machine learning algorithms," *Exploration of Digital Health Technologies*, pp. 30–45, 2024, doi: 10.37349/edht.2024.00009.
[7] B. Hartman, R. Owen, and Z. Gibbs, "Predicting high-cost health insurance members through boosted trees and oversampling: an application using the HCCI database," *North American Actuarial Journal*, pp. 1–9, 2020, doi: 10.1080/10920277.2020.1754242.
[8] D. A. Vallero, "Predicting risks in an increasingly complex world," *Environmental Systems Science*, pp. 89–133, 2021, doi: 10.1016/b978-0-12-821953-9.00006-4.
[9] U. W. de Ruijter *et al.*, "Prediction models for future high-need high-cost healthcare use: a systematic review," *Journal of General Internal Medicine*, vol. 37, no. 7, pp. 1763–1770, 2022, doi: 10.1007/s11606-021-07333-z.
[10] Z. Li *et al.*, "Developing a model to predict high health care utilization among patients in a New York City safety net system," *Medical Care*, vol. 61, no. 2, pp. 102–108, 2023, doi: 10.1097/MLR.0000000000001807.
[11] E. M. Alotaibi, "Risk assessment using predictive analytics," *International Journal of Professional Business Review*, vol. 8, no. 5, 2023, doi: 10.26668/businessreview/2023.v8i5.1723.

[12] S. T. Moturu, W. G. Johnson, and H. Liu, "Predictive risk modelling for forecasting high-cost patients: a real-world application using medicaid data," *International Journal of Biomedical Engineering and Technology*, vol. 3, no. 1–2, pp. 114–132, 2010, doi: 10.1504/IJBET.2010.029654.
[13] M. Mashasha, P. Mutize, and F. Mazunga, "Distribution and pattern of an insurance health claim system: a time series approach," *Tanzania Journal of Science*, vol. 48, no. 1, pp. 13–21, 2022, doi: 10.4314/tjs.v48i1.2.
[14] S. A. Aloyuni, "A systematic review on machine learning and deep learning based predictive models for health informatics," *Journal of Pharmaceutical Research International*, pp. 183–194, 2021, doi: 10.9734/jpri/2021/v33i47b33112.
[15] E. E. Agu, A. O. Abhulimen, A. N. O.-Osafiele, O. S. Osundare, I. A. Adeniran, and C. P. Efunniyi, "Utilizing AI-driven predictive analytics to reduce credit risk and enhance financial inclusion," *International Journal of Frontline Research in Multidisciplinary Studies*, vol. 3, no. 2, pp. 20–29, 2024, doi: 10.56355/ijfrms.2024.3.2.0026.
[16] H. Abdulazeem, S. Whitelaw, G. Schauberger, and S. J. Klug, "A systematic review of clinical health conditions predicted by machine learning diagnostic and prognostic models trained or validated using real-world primary health care data," *PLOS ONE*, vol. 18, no. 9, p. e0274276, Sep. 2023, doi: 10.1371/journal.pone.0274276.
[17] G. Tse *et al.*, "Healthcare big data in Hong Kong: development and implementation of artificial intelligence-enhanced predictive models for risk stratification," *Current Problems in Cardiology*, vol. 49, no. 1, 2024, doi: 10.1016/j.cpcardiol.2023.102168.
[18] P. S. Kewalchand, "AI in healthcare," *International Journal of Advanced Research in Science, Communication and Technology*, vol. 4, no. 2, pp. 548–554, 2024. [Online] Available: https://ijarsct.co.in/Paper15285.pdf
[19] E. Nabrawi and A. Alanazi, "Fraud detection in healthcare insurance claims using machine learning," *Risks*, vol. 11, no. 9, 2023, doi: 10.3390/risks11090160.
[20] C. Crowley, J. Perloff, A. Stuck, and R. Mechanic, "Challenges in predicting future high-cost patients for care management interventions," *BMC Health Services Research*, vol. 23, no. 1, 2023, doi: 10.1186/s12913-023-09957-9.
[21] D. Mwembe, B. Jones, W. Chagwiza, and S. Ngwenya, "A statistical analysis of time to a claim: a case of Zimbabwe's health insurance clientele," *International Journal of Applied Business and Management Sciences*, vol. 3, no. 2, pp. 267–288, 2022, doi: 10.47509/IJABMS.2022.v03i02.07.
[22] C. A. Ardagna, P. Ceravolo, and E. Damiani, "Big data analytics as-a-service: issues and challenges," in *2016 IEEE International Conference on Big Data (Big Data)*, IEEE, Dec. 2016, pp. 3638–3644. doi: 10.1109/BigData.2016.7841029.
[23] S. Akter, Y. K. Dwivedi, K. Biswas, K. Michael, R. J. Bandara, and S. Sajib, "Addressing algorithmic bias in AI-driven customer management," *Journal of Global Information Management*, vol. 29, no. 6, 2021, doi: 10.4018/JGIM.20211101.oa3.
[24] I. Kapungu, E. Evarest, N. Shaban, and A. J. Mwakisisile, "Modelling and forecasting claim payments of Tanzania national health insurance fund," *Tanzania Journal of Science*, vol. 49, no. 4, pp. 911–920, Oct. 2023, doi: 10.4314/tjs.v49i4.12.
[25] M. Nalluri, M. Pentela, and N. R. Eluri, "A scalable tree boosting system: XG boost," *International Journal of Research Studies in Science, Engineering and Technology*, vol. 7, no. 12, pp. 36–51, 2020, doi: 10.22259/2349-476X.0712005.

BIOGRAPHIES OF AUTHORS

Lenias Zhou is an MTech Student in Cloud Computing at Harare Institute of Technology (HIT), Zimbabwe, a holder of a Bachelor of Science Honors Degree in Information Systems from the Women's University in Africa, and a holder of a Full Technological Diploma in Telecommunications and Electronics Engineering from the City and Guild Institute, London. He is a Certified Information Systems Auditor (CISA) with 25 years of experience in technical operations and technical audits. He has led several projects in telecommunications, IT, and audit. His broad research interests cover topics relating to AI, cybersecurity, IoT, cloud computing, and network engineering. He can be contacted at email: zhoulenias@gmail.com.

Mainford Mutandavari is a PhD Scholar at SRMIST University, India, and a Lecturer and Postgraduate Studies Coordinator at the Harare Institute of Technology (HIT), Zimbabwe. With advanced degrees in Computer Science and Strategy and Innovation, his research spans data analytics, cybersecurity, IoT, AI, and cloud computing. He is a member of HIT's Cybersecurity and AI research groups and actively contributes to national ICT standards through the Standards Association of Zimbabwe. Minford has published widely on topics such as data loss prevention systems, digital learning infrastructure, and e-health security. His work bridges academic research with industry applications, focusing on practical digital solutions for education, telecommunications, and healthcare in Zimbabwe. He is also involved in curriculum development, postgraduate supervision, and building academic-industry partnerships. He can be contacted at email: mmtandavari@gmail.com.

Lucia Matondora is a graduate with a BTech and MTech in Software Engineering from Harare Institute of Technology. She worked in the industry for a period of a year as a System Support Officer and then came back to Harare Institute of Technology and started her Master of Technology in Software Engineering in the Information Security and Assurance department. She aimed to improve her knowledge as well as her expertise in the field of software engineering. She loves mentoring and guiding students, especially the girl child, to empower them. She is a tech enthusiast and loves researching the impact of new technologies in education systems. She can be contacted at email: lmatondora@hit.ac.zw.