International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(10s) 2026 pp. 738-749
https://doi.org/10.70917/ijcisim-2026-3646
Article

MACHINE LEARNING CLASSIFICATION OF COMMUTING ACCIDENT CLAIM OUTCOMES

Mohd Zaki Awang Chek[1]*, Hilmi Aziz Bukhori[2]

[1] Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Perak Branch, Tapah Campus, 35400 Tapah, Perak, Malaysia
[2] Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Malang 65145, Indonesia
*Corresponding author: mohdz220@uitm.edu.my
Second-author email: hilmibukhori@ub.ac.id

Abstract: Background: Commuting accident claims impose substantial occupational, social, and administrative burdens on social security systems. Despite the increasing availability of administrative claims databases, evidence on machine-learning approaches for predicting commuting accident claim outcomes remains limited, particularly in the context of national social insurance programmes.
Methods: This study developed and compared five supervised machine-learning algorithms for classifying commuting accident claim outcomes using 37,828 commuting accident claim records obtained from the Social Security Organisation (SOCSO), Malaysia. The analytical dataset comprised 57 demographic, accident, injury, and administrative variables, with approved claims representing 61.5% of all observations. The dataset was randomly partitioned into 70% training (n = 26,480) and 30% validation (n = 11,348) subsets using SAS Enterprise Miner. Five classification models were evaluated: Decision Tree, Interactive Decision Tree, Gradient Boosting, Logistic Regression, and Neural Network. Model performance was assessed using recall, receiver operating characteristic area under the curve (ROC-AUC), misclassification rate, and average squared error.
Results: Gradient Boosting achieved the highest predictive performance, producing a validation ROC-AUC of 0.998 and the lowest validation misclassification rate (0.0121). The Interactive Decision Tree demonstrated the greatest model stability, with minimal differences between the training and validation datasets in misclassification rate (0.000009) and average squared error (0.000022), while providing transparent and interpretable decision rules. Variable importance analysis consistently identified Reason_Code as the strongest predictor across several models. However, because this variable may be generated during the claim adjudication process, its inclusion requires careful verification to exclude potential target leakage and ensure unbiased model evaluation.
Conclusions: The findings demonstrate the potential of machine-learning techniques for accurately classifying commuting accident claim outcomes using large-scale social security administrative data. Among the evaluated models, Gradient Boosting provided the highest predictive accuracy, whereas the Interactive Decision Tree offered a favourable balance between predictive performance and interpretability for operational decision support. Nevertheless, future studies should employ leakage-controlled feature selection, independent or temporal validation datasets, calibration assessment, and fairness analyses before these models are considered for implementation in operational claim management systems.
Keywords: commuting accidents; claim outcome prediction; machine learning; gradient boosting; decision tree; occupational safety; social security; insurance analytics; Malaysia.

1. Introduction

Commuting accidents occur while workers travel between their residence and workplace or another recognized location connected with employment. Although these incidents take place outside the physical workplace, they can produce occupational injuries, temporary or permanent disability, lost earnings, family disruption, and substantial social-insurance expenditure. In Malaysia, commuting accidents are therefore relevant not only to road safety but also to occupational protection and the administration of statutory employment-injury benefits (Aziz & Yusof, 2015; Brati et al., 2025; Lutfi Abas et al., 2013; Sahak et al., 2020; Selamat & Surienty, 2015).

Malaysia's Social Security Organisation (SOCSO) maintains administrative records on employment-related injuries and claims. These records contain demographic, accident, injury, medical, and claim-related information that may support systematic classification of claim outcomes. Conventional manual assessment remains essential because legal and documentary requirements must be considered; however, carefully validated predictive models may assist administrative triage by identifying records that warrant prioritization or additional review (Baharudin et al., 2013; Chek et al., 2019).

Previous Malaysian research has described the circumstances and consequences of commuting accidents, including the characteristics of injured workers and the legal responsibilities surrounding employment-related travel. Nevertheless, fewer studies have examined the administrative classification of commuting accident claim status using large social-security datasets. Existing predictive studies also frequently emphasize discrimination alone, while giving less attention to predictor timing, data leakage, calibration, independent validation, interpretability, and the operational conditions under which a model would be used (Khairuddin et al., 2024; Selamat & Surienty, 2015; Zuwairy et al., 2020).

The present study therefore compared a default Decision Tree, an Interactive Decision Tree, Gradient Boosting, Logistic Regression, and a Neural Network for classifying SOCSO commuting accident claim status. The study focused on three questions: (1) how well did the five models classify claim status under the reported internal validation design; (2) which variables contributed most strongly to classification; and (3) what trade-off emerged between predictive performance and interpretability? (Basheer & Hajmeer, 2000; Breiman, 2001; B. Chen & others, 2026; T. Chen & Guestrin, 2016; LaValley, 2008).

The principal contribution is empirical and contextual rather than algorithmic. The study applies established statistical and machine-learning methods to a large administrative dataset and compares their reported performance in a Malaysian social-security setting. It also highlights a central implementation issue: a model intended for prospective claim triage must use only variables available at the time the prediction is made (Adamu-Fika et al., 2023; Peng & So, 2002; Zuwairy et al., 2020).

2. Related Literature

2.1 Commuting accidents and social protection

Commuting accidents represent an intersection between transport safety, occupational health, and social protection. Malaysian studies have examined commuting-injury patterns, survivor characteristics, legal responsibilities, and the broader burden placed on workers and their. These studies establish the importance of the problem but do not, by themselves, demonstrate that administrative claim outcomes can be predicted reliably or used to prevent accidents (Khairuddin et al., 2024; Selamat & Surienty, 2015; Zuwairy et al., 2020) .

Claim-status modelling differs from accident-risk or accident-severity modelling. A claim-status classifier estimates an administrative outcome, whereas an accident-risk model estimates the probability that an accident will occur and a severity model estimates injury consequences. These purposes require different predictors, validation strategies, interpretations, and policy claims. The present manuscript is therefore framed specifically as a claim-status classification study (B. Chen & others, 2026; James et al., 2023; Khairuddin et al., 2024; Koster et al., 2021; LaValley, 2008).

2.2 Predictive modelling for administrative claims

Administrative claim records are attractive for predictive modelling because they are structured, routinely collected, and often available at scale. However, they also create methodological risks. Variables may be entered or updated at different stages of adjudication, and fields recorded after a decision may encode the outcome directly or indirectly. Accordingly, predictive validity depends not only on model accuracy but also on a clearly defined prediction time and a documented temporal eligibility rule for every predictor (James et al., 2023; Selamat & Surienty, 2015).

Interpretable models may be particularly valuable in administrative settings because staff must understand why a record is assigned a particular risk or classification. Decision trees provide explicit split rules, while logistic regression provides coefficients and odds ratios when model assumptions are reasonably satisfied. Ensemble and neural models may offer stronger predictive performance but often require additional interpretation and calibration analyses before deployment (Khairuddin et al., 2024; Selamat & Surienty, 2015).

2.3 Evaluated algorithms

Decision trees recursively partition observations according to predictor values and generate transparent classification rules. Interactive tree development may permit analyst-guided modification of splits, although such intervention must be documented and evaluated on data not used to guide the modifications. Gradient Boosting combines a sequence of weak learners, with each learner attempting to reduce errors from the preceding ensemble. Its flexibility can capture nonlinear effects and interactions, but performance depends on settings such as tree depth, number of trees, learning rate, subsampling, and stopping criteria (Su & Bai, 2020).

Logistic Regression estimates the probability of a binary outcome through the logistic function. It is readily interpretable, but sparse categories, multicollinearity, missing data, and complete or quasi-complete separation can destabilize coefficients and produce extreme odds ratios (LaValley, 2008; Peng & So, 2002).

Artificial neural networks can model nonlinear relationships by transforming inputs through hidden units and activation functions. Their replication requires complete reporting of architecture, optimization, regularization, stopping, and preprocessing decisions (Basheer & Hajmeer, 2000).

3. Methods

3.1 Study design and data source

This retrospective predictive modelling study utilised administrative commuting accident claim records obtained from the **Social Security Organisation (SOCSO), Malaysia**. The analytical dataset last three years (2023 – 2025) comprised **37,828 commuting accident claim records** all around Malaysia described by **57 variables**, including demographic characteristics, accident circumstances, injury characteristics, employer information, administrative processing variables, and claim outcomes. The unit of analysis was an individual commuting accident claim.

The dataset was extracted from SOCSO's administrative claims database and represents commuting accident claims processed during the study period. Prior to analysis, the dataset was reviewed for completeness and suitability for predictive modelling.

3.2 Outcome and predictors

The outcome variable was claim status, representing the final administrative decision for each commuting accident claim. The original dataset contained multiple claim-status categories, with approved claims accounting for 23,265 (61.50%) of the 37,828 records. For supervised classification, the claim-status variable was transformed into a binary target variable in SAS Enterprise Miner, where approved claims constituted one class and all remaining claim outcomes constituted the comparison class. Candidate predictor variables were selected from information routinely available in the administrative database and included:

- demographic characteristics (e.g., age, gender, ethnicity);
- accident characteristics (e.g., accident timing and accident cause);
- injury characteristics (e.g., injury type and injury location); and
- administrative variables (e.g., Reason_Code, claim amount, number of medical certificates, employer information, and claim-processing variables).

Variables exhibiting substantial missingness or recorded after claim adjudication were evaluated carefully because they may introduce information leakage into predictive models.

3.3 Variable-role and leakage audit

A total of **57 candidate variables** were initially available for modelling. Variables were assigned analytical roles in SAS Enterprise Miner as **Input**, **Target**, or **Rejected**, depending on their relevance to model development.

Because the objective was to predict claim outcomes, predictor eligibility was assessed according to information availability before claim adjudication. Demographic variables (age, gender, ethnicity), accident descriptors, and injury characteristics were considered appropriate predictors because they are generally available during the initial claim submission. Conversely, administrative variables generated during claim processing including **Reason_Code**, **claim amount**, **payment information**, and **number of medical certificates** may only become available after adjudication and therefore present a potential risk of target leakage. To ensure valid model development, only variables available at the intended prediction time should be included in prospective operational models. Variables generated during the adjudication process should either be excluded or their temporal availability clearly documented (Jamal, Abdul Ghafar, Ismail, & Awang Chek, 2018; Woodfield, 2010).

Variable	**Reported role**	**Likely type**	**Availability before decision**	**Risk**	**Required action**
Claim status	Target	Binary	Outcome	—	Define positive and negative classes
Age group	Input	Nominal/ordinal	Likely available	Low	Confirm coding and missingness
Gender	Input	Nominal	Likely available	Fairness	Report subgroup performance
Ethnicity	Input	Nominal	Likely available	Fairness	Justify use and assess bias
Accident time	Input	Nominal/time	Likely available	Low	Define categories
Accident cause	Input	Nominal	Verify	Moderate	Confirm when finalized
Injury type/location	Input	Nominal	Verify	Moderate	Confirm initial-report availability
Reason_Code	Input	Nominal	Unknown	Critical	Exclude unless prospectively available
Claim amount	Input	Interval	Unknown	Critical	Likely exclude from primary model
Medical certificates	Input	Interval	Unknown	High	Define count as of prediction time

3.4 Data preparation

Data preparation was conducted using SAS Enterprise Miner. Variable roles were assigned according to the analytical objective, and categorical variables were encoded automatically by the software during model training.

The dataset was randomly partitioned into a training dataset (70%; n = 26,480) and a validation dataset (30%; n = 11,348) using the Data Partition node in SAS Enterprise Miner. Model development was performed exclusively on the training data, while model performance was evaluated using the independent validation dataset.

Exploratory analysis showed that most demographic and accident-related variables contained minimal missing values. However, several administrative variables exhibited substantial missingness, including Profession Code (100%), Date of Death (96.02%), Unifile (92.98%), Number of Medical Certificates (10.49%), and Claim Amount (9.29%). Variables with excessive missingness or limited analytical value were considered during variable selection to minimise bias and improve model robustness.

Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
ACCIDENT_TIN	Input	Nominal	No		No	.	.
AGE_GROUP	Input	Nominal	No		No	.	.
Amaun_Bayara	Input	Interval	No		No	.	.
Bil_MC	Input	Nominal	No		No	.	.
CAUSE_OF_AC	Input	Nominal	No		No	.	.
CLAIM_STATU	Target	Binary	No		No	.	.
Ethnic_Code	Input	Nominal	No		No	.	.
File_No	ID	Nominal	No		No	.	.
Gender_Code	Input	Nominal	No		No	.	.
Kadar_Harian_	Input	Interval	No		No	.	.
LOCATION_OF	Input	Nominal	No		No	.	.
Reason_Code	Input	Nominal	No		No	.	.
TYPE_OF_INJU	Input	Nominal	No		No	.	.

Figure 1 illustrates the variable-role assignments implemented in SAS Enterprise Miner.

3.5 Model development

Five supervised machine-learning algorithms were developed and compared using identical training and validation datasets (Brati et al., 2025).

3.5.1 Decision Tree

A default Decision Tree classifier was constructed using recursive binary partitioning to classify commuting accident claim outcomes. The algorithm selected predictor variables based on impurity reduction and generated hierarchical decision rules (De Reyck et al., 2008).

3.5.2 Interactive Decision Tree

An Interactive Decision Tree was developed to improve model interpretability through analyst-guided refinement of the tree structure. This approach enabled domain knowledge to be incorporated into node selection while maintaining transparent decision rules suitable for operational interpretation (Kose et al., 2015).

3.5.3 Gradient Boosting

Gradient Boosting was implemented as an ensemble learning algorithm that sequentially combined multiple weak decision trees to minimise prediction error. The boosting procedure enhanced predictive performance by iteratively correcting classification errors from previous trees (Su & Bai, 2020).

3.5.4 Logistic Regression

Binary Logistic Regression was employed as a statistical baseline classifier. Variable selection was performed using forward stepwise selection within SAS Enterprise Miner to estimate the probability of claim approval based on the selected predictor variables (Jamal, Abdul Ghafar, Ismail, Awang Chek, et al., 2018).

3.5.5 Neural Network

A feed-forward Neural Network was developed to capture nonlinear relationships between predictor variables and claim outcomes. The network architecture and model parameters were determined using SAS Enterprise Miner, and model performance was compared with the other machine-learning approaches (Basheer & Hajmeer, 2000).

3.6 Validation and model selection

Model performance was evaluated using the independent **30% validation dataset**. The study did not employ an external validation dataset, repeated cross-validation, or temporal validation; therefore, all reported results represent **internal validation**. Model selection considered both predictive accuracy and interpretability. Although **Gradient Boosting** achieved the highest predictive performance, the **Interactive Decision Tree** provided transparent decision rules with minimal differences between training and validation performance, making it a potentially useful decision-support tool for operational implementation (Lundberg & Lee, 2017).

3.7 Evaluation measures

Model performance was evaluated using several complementary classification metrics, including Receiver Operating Characteristic Area Under the Curve (ROC-AUC), recall (sensitivity), misclassification rate, and average squared error. ROC-AUC was used to assess the models' ability to discriminate between approved and non-approved claims across all decision thresholds. Recall measured the proportion of correctly identified positive cases, whereas the misclassification rate quantified the overall proportion of incorrectly classified observations. Average squared error provided an additional measure of prediction accuracy. Comparing multiple evaluation metrics enabled a comprehensive assessment of predictive performance while balancing classification accuracy, model stability, and interpretability (Gustman & Steinmeier, 1999).

3.8 Ethical, privacy, and fairness considerations

The study utilised anonymised administrative claim records obtained from the Social Security Organisation (SOCSO), Malaysia. All analyses were conducted using de-identified data to protect claimant confidentiality and privacy. Because demographic variables such as age, gender, and ethnicity were included as predictors, model performance should be assessed across population subgroups before operational deployment to minimise potential algorithmic bias. Furthermore, predictor importance should not be interpreted as evidence of causal relationships but rather as statistical associations within the observed administrative data.

4. Results

4.1 Dataset overview

The study analysed 37,828 commuting accident claim records obtained from the Social Security Organisation (SOCSO), Malaysia, comprising 57 variables describing demographic characteristics, accident circumstances, injury characteristics, employer information, administrative processing, and claim outcomes. The primary outcome variable was claim status, while the predictor variables included demographic information (e.g., age, gender, and ethnicity), accident characteristics (e.g., accident time and accident cause), injury-related variables (e.g., injury type and injury location), and administrative variables (e.g., Reason_Code, claim amount, and number of medical certificates).

The study population consisted predominantly of male claimants (29,757; 78.66%), whereas female claimants accounted for 8,071 (21.34%) of all records. Most accidents occurred during commuting to or from work (33,538; 88.65%), followed by rest-break travel (2,118; 5.60%) and other commuting circumstances (2,172; 5.74%). Fractures were the most frequently reported injury type (14,883; 39.34%), while striking against moving objects represented the most common accident mechanism (14,633; 38.68%). Among the recorded claim outcomes, 23,265 (61.50%) claims were classified as approved.

Before model development, the dataset was randomly partitioned using the Data Partition node in SAS Enterprise Miner, allocating 26,480 records (70%) to the training dataset for model development and 11,348 records (30%) to the validation dataset for internal model evaluation. Most demographic and accident-related variables contained minimal missing data; however, several administrative variables, including Profession Code, Date of Death, and Unifile, exhibited substantial missingness and were treated appropriately during data preprocessing. A comprehensive summary of the study population, outcome distribution, and data completeness is presented in table below.

Domain	Characteristic	n	%
Dataset profile	Total commuting accident claim records	37,828	100.00
	Total variables	57	—
Gender	Male	29,757	78.66
	Female	8,071	21.34
Ethnicity	Malay	6,727	17.79
	Chinese	2,747	7.26
	Indian	3,465	9.16
	Other ethnic groups	349	0.92
	Not recorded	24,540	64.87
Accident circumstance	Commuting to or from work	33,538	88.65
	During a rest or meal break	2,118	5.60
	Other commuting circumstances	2,172	5.74
Most frequently recorded accident causes	Striking against moving objects	14,633	38.68
	Falls from a height or into a depth	6,584	17.41
	Falls on the same level	5,461	14.44
	Other accident types	2,857	7.55
	Struck by moving objects	2,654	7.02
Most frequently recorded injury types	Fractures	14,883	39.34
	Other wounds	12,830	33.92
	Multiple injuries	2,494	6.59
	Other or unspecified injuries	1,483	3.92
	Internal injuries or concussion	1,362	3.60
Most frequently recorded injury locations	Multiple body locations	5,284	13.97

	Hand	4,313	11.40
	Lower leg	3,735	9.87
	Shoulder	2,638	6.97
	Foot	2,437	6.44
Claim-status distribution	Approved	23,265	61.50
	Claim type changed	6,574	17.38
	Rejected	3,152	8.33
	Exported to PPT	1,935	5.12
	Closed	1,320	3.49
	Other recorded statuses	1,582	4.18
Data partition	Training set	26,480	70.00
	Validation set	11,348	30.00
	Independent test set	0	0.00
Selected variables with missing values	Profession Code	37,828	100.00
	Date of Death	36,323	96.02
	Unifile	35,174	92.98
	Number of medical certificates	3,968	10.49
	Claim or payment amount	3,514	9.29
	Claim completion date	379	1.00
	Age	11	0.03

4.2 Tree-based model interpretation

Reason_Code was the first split in both reported tree structures, indicating that it carried substantial information about claim status in the analysed data. In the default tree, claim amount appeared as a subsequent split, while the interactive tree additionally used the number of medical certificates. Because these variables may arise during claim processing, the observed predictive strength cannot be interpreted as prospective usefulness until their timing has been verified.

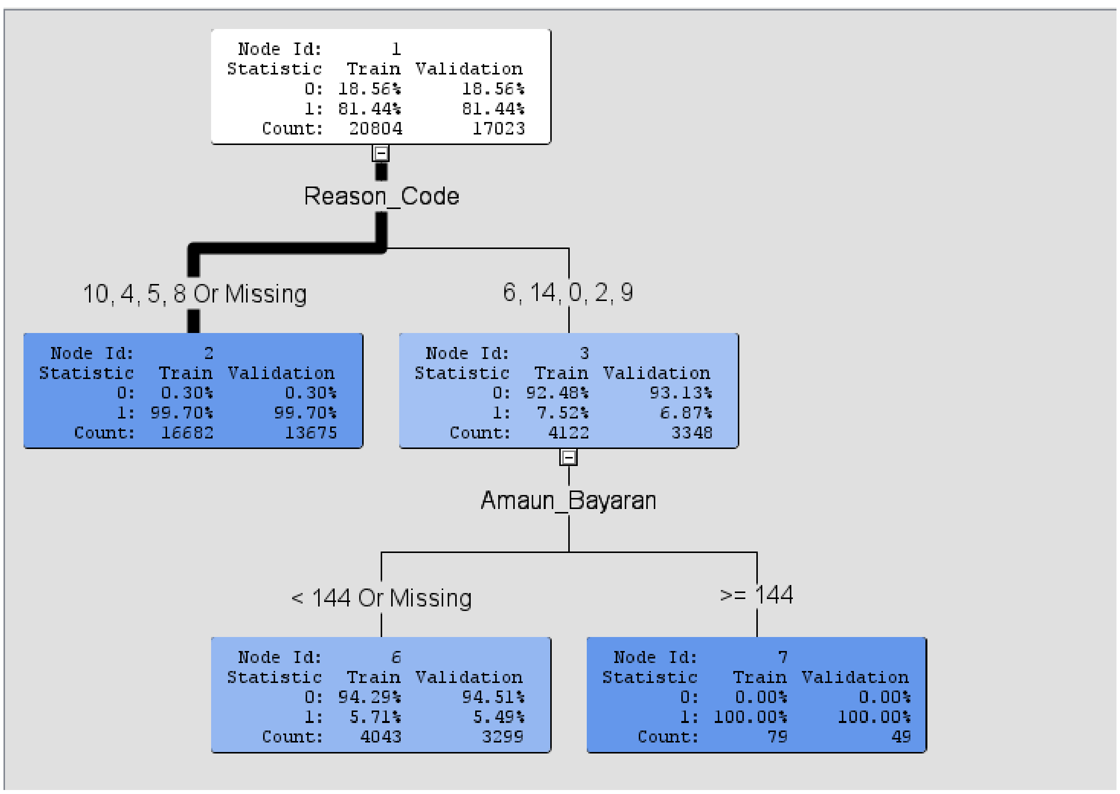

Figure 2. Reported default Decision Tree. Reason_Code was the root split and claim amount was used subsequently.

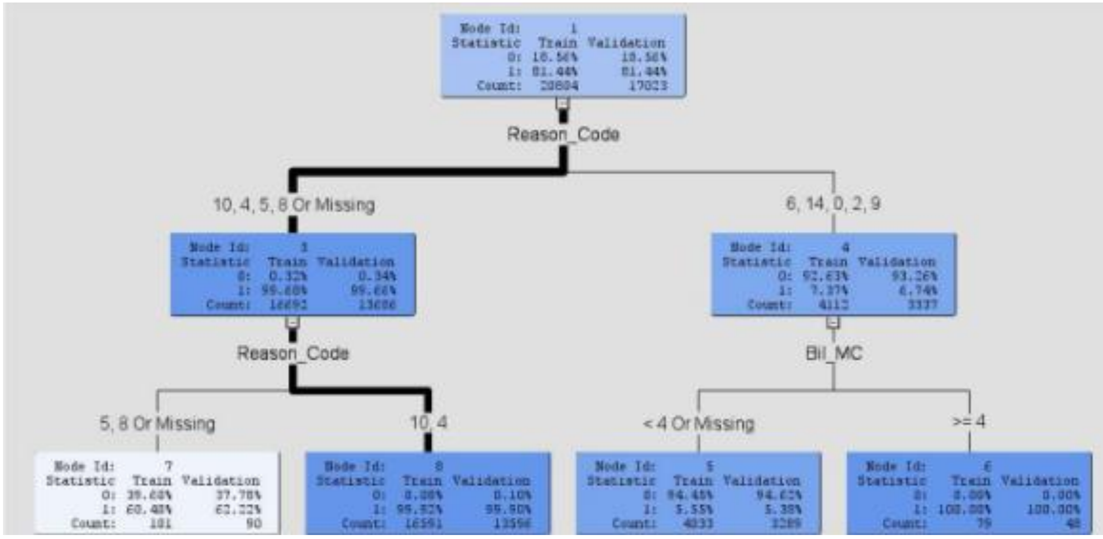

Figure 3. Reported Interactive Decision Tree. Reason_Code and the number of medical certificates were used as major splits.

4.3 Comparative performance

Table 1 reproduces the most internally coherent performance table from the source manuscript. Gradient Boosting had the lowest validation misclassification rate (0.012101) and the highest reported validation ROC index (0.998). The Interactive Decision Tree had the smallest training-validation gaps in misclassification rate (0.000009) and average squared error (0.000022). These findings support different conclusions: Gradient Boosting was the strongest reported predictive model, whereas the Interactive Decision Tree was the most transparent model with the smallest reported error gaps.

Model	**Misclassification rate**			**ROC index**			**Average squared error**		
	Train	**Validation**	**Gap**	**Train**	**Validation**	**Gap**	**Train**	**Validation**	**Gap**
Gradient Boosting	0.011536	0.012101	0.000565	0.999	0.998	0.001	0.011821	0.013178	0.001357
Decision Tree	0.013507	0.013041	0.000466	0.987	0.987	0.000	0.012865	0.012452	0.000413
Interactive Decision Tree	0.013267	0.013276	0.000009	0.993	0.993	0.000	0.011728	0.011750	0.000022
Neural Network	0.016199	0.015332	0.000867	0.994	0.994	0.000	0.013211	0.012847	0.000364
Logistic Regression	0.016872	0.015920	0.000952	0.996	0.996	0.000	0.012674	0.012597	0.000077

Table 1. Reported training and validation performance from the source manuscript. Values require verification against the original SAS output because other source tables contain conflicting ROC and average-squared-error values.

The reported validation accuracy corresponding to the Gradient Boosting misclassification rate was approximately 98.79% (1 - 0.012101). The same relationship should be checked for every model before final submission. No confidence intervals or statistical comparisons were reported, so small differences between models should not be interpreted as conclusive.

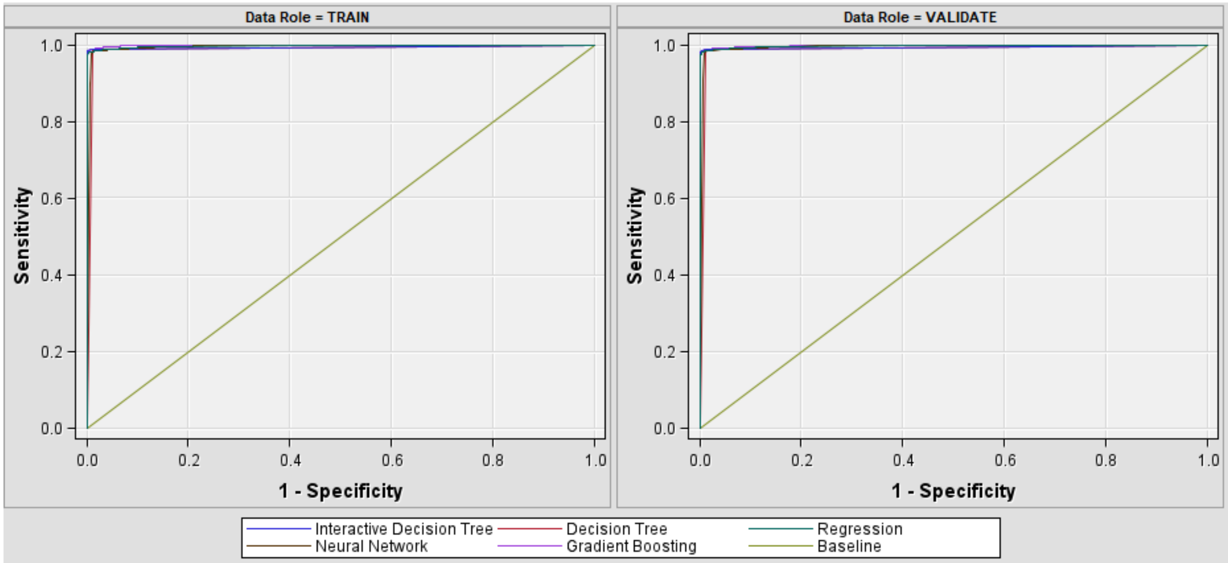

Figure 4. Reported ROC curves for the training and validation partitions. The plotted values should be regenerated from the finalized leakage-controlled analysis.

4.4 Logistic Regression findings

The source manuscript reported very large odds ratios for selected Reason_Code categories, including 461.919 and 999.000. Such estimates may indicate strong administrative association, sparse categories, complete or quasi-complete separation, or target leakage. They should not be interpreted substantively until category frequencies, coefficients, standard errors, confidence intervals, and predictor timing have been examined.

4.5 Neural Network findings

The source manuscript identified the configuration described as having five hidden layers as the preferred neural network because it had the smallest reported training-validation average-squared-error gap. However, the architecture and training procedure were not sufficiently documented to support replication, and the selected terminology requires verification.

5. Discussion

5.1 Principal findings

The comparative analysis suggests that Gradient Boosting achieved the strongest reported internal predictive performance, whereas the Interactive Decision Tree offered transparent rules and the smallest reported differences between training and validation error. This pattern illustrates a familiar operational trade-off: more flexible ensemble models may improve discrimination, while tree models may be easier for claims personnel to inspect and communicate (Su & Bai, 2020).

Reason_Code dominated several models. This result may reflect genuine information about the type or basis of a claim, but it may also reflect the administrative workflow through which claim decisions are recorded. Without a clear timestamp and definition, the variable cannot be treated as an independent prospective predictor.

5.2 Comparison and interpretation

The findings are broadly consistent with the capacity of tree-based and ensemble models to identify nonlinear interactions and high-order partitions in structured administrative data. However, the very high internal performance reported across all five models is unusual enough to require careful scrutiny. Potential explanations include an easily separable outcome, strong administrative codes, a highly imbalanced target, repeated use of validation data, or leakage from variables recorded during adjudication (Selamat & Surienty, 2015).

The Interactive Decision Tree should not be described as the overall best model solely because its training and validation errors were close. A small gap may indicate stability, but it can also occur when both estimates are biased or when the model underfits. Operational selection should instead consider validated discrimination, calibration, uncertainty, interpretability, fairness, and the consequences of false approvals and false rejections (Sahak et al., 2020).

5.3 Practical implications

After appropriate validation, a claim-status model may support administrative triage by prioritizing records for routine processing, additional documentation, or expert review. It should not replace statutory adjudication or be used as the sole basis for approving or rejecting a claim. Human oversight, transparent decision rules, audit trails, and procedures for contesting model-assisted decisions would be essential. The present analysis does not demonstrate that machine learning prevents commuting accidents, detects fraud, reduces costs, or shortens processing time. Those outcomes would require prospective implementation studies with predefined operational and safety endpoints (Khairuddin et al., 2024; Selamat & Surienty, 2015).

5.4 Responsible use and fairness

A model trained on historical administrative decisions may reproduce past inconsistencies or disparities. Before implementation, SOCSO or any comparable organization should examine false-positive and false-negative rates across relevant demographic and occupational groups, assess whether sensitive variables are necessary, and evaluate whether alternative models produce materially different distributions of error.

5.5 Strengths and limitations

The study's principal strength is the use of a large real-world administrative dataset and the comparison of statistical, tree-based, ensemble, and neural approaches within a common analytical environment. The tree structures also provide a preliminary view of the variables driving classification. Several limitations materially restrict interpretation. First, the retrospective administrative design does not establish causal relationships. Second, the study period, inclusion criteria, missing-data procedures, class distribution, and data-partition strategy were not sufficiently reported. Third, Reason_Code, claim amount, and medical-certificate counts may introduce target leakage if they were recorded after the intended prediction point. Fourth, the analysis lacked an independent test set, repeated cross-validation, calibration measures, uncertainty intervals, fairness assessment, temporal validation, and external validation. Fifth, model hyperparameters and neural-network architecture were incompletely documented. Finally, the source manuscript contained conflicting performance values across tables, requiring verification against the original SAS outputs (Basheer & Hajmeer, 2000; Khairuddin et al., 2024; SAS, n.d.).

6. Recommended Further Analysis

The primary model should be retrained using only predictors demonstrably available at claim intake or another prespecified decision point. A secondary exploratory model may include later administrative variables, but it should be labelled clearly and should not be presented as prospectively deployable. The recommended evaluation design is: (1) reserve an untouched temporal or random test set; (2) conduct hyperparameter tuning within repeated stratified cross-validation on the development data; (3) report ROC AUC, precision-recall AUC, sensitivity, specificity, precision, F1-score, confusion matrices, Brier score, calibration intercept and slope, and 95% confidence intervals; (4) compare models using paired methods where appropriate; and (5) conduct subgroup and sensitivity analyses.

7. Conclusion

This study compared five methods for classifying commuting accident claim status using 37,828 SOCSO administrative records. Gradient Boosting achieved the strongest reported internal predictive performance, whereas the Interactive Decision Tree provided more transparent rules and the smallest reported training-validation error gaps. Reason_Code was the dominant predictor across several models, but its administrative timing must be established before the result can be interpreted as prospective predictive evidence. The present findings demonstrate the analytical potential of SOCSO data but do not yet justify operational deployment or conclusions about accident prevention, fraud detection, processing efficiency, or policy effectiveness. Future work should apply a prespecified prediction time, remove leakage-prone variables, document reproducible model tuning, evaluate calibration and fairness, quantify uncertainty, and validate the final model on independent temporal or external data. Following these steps, the framework may provide a defensible foundation for transparent, human-supervised claim-triage support (Zuwairy et al., 2020).

Acknowledgements

The authors gratefully acknowledge the Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM) Perak Branch, Tapah Campus, Malaysia, for providing the academic environment and institutional support for this research. The authors also express their sincere appreciation to the Social Security Organisation (SOCSO) Malaysia for granting access to the commuting accident claims data used in this study. In addition, the authors acknowledge the Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Indonesia, for fostering international academic collaboration through the Visiting Lecturer Programme 2026, which facilitated scholarly exchange and joint research activities that contributed to the development of this manuscript.

Declarations

Funding

This research received direct external research funding. The study was conducted within the framework of the Universitas Brawijaya Visiting Lecturer Programme 2026, an institutionally supported international academic collaboration initiative between Universitas Brawijaya, Indonesia, and Universiti Teknologi MARA (UiTM), Malaysia. The programme facilitated academic exchange, collaborative research, and manuscript development; however, no specific research grant or financial support was provided for this study.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data availability

The data supporting this study were obtained from Malaysia's Social Security Organisation and may contain confidential administrative claim information. Access is subject to organizational authorization and applicable data-protection requirements. [Confirm whether anonymized metadata, code, or a synthetic dataset can be shared.]

Author contributions

Mohd Zaki Awang Chek: Conceptualization, Methodology, Formal analysis, Investigation, Writing - original draft, and Project administration. Hilmi Aziz Bukhori: Validation, Writing - review and editing, and Academic collaboration.

References

1. Adamu-Fika, F., Madaki, D. B., Baba-Onoja, A. E., Ramalan, A. T., Mohammed, A. T., & Bature, K. S. (2023). Modelled Machine Learning Algorithms to Predict Students' Academic Performance in Tertiary Institutions. Advances in Multidisciplinary Scientific Research Conference Proceedings, 418–427.
2. Aziz, N. H. A., & Yusof, A. A. (2015). The Employer's Duties and Liabilities in Commuting Accidents in Malaysia: Law and Management. Procedia - Social and Behavioral Sciences, 211, 796–802. https://doi.org/10.1016/j.sbspro.2015.11.170
3. Baharudin, M. R., Siang, L. C., Yeng, S. W., & Nazmin, H. (2013). Effectiveness of SOCSO's accident prevention program in Klang Valley, 2012.
4. Basheer, I. A., & Hajmeer, M. (2000). Artificial Neural Networks: Fundamentals, Computing, Design, and Application. Journal of Microbiological Methods, 43(1), 3–31. https://doi.org/10.1016/S0167-7012(00)00201-3
5. Brati, E., Braimllari, A., & Gjeçi, A. (2025). Machine Learning Applications for Predicting High-Cost Claims Using Insurance Data. Data, 10(6), 90. https://doi.org/10.3390/data10060090
6. Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32.
7. Chek, M. Z. A. M. Z. A., Ismail, I. L. I. L., & Jamal, N. F. N. F. (2019). Descriptive research on SOCSO's Invalidity Pension Scheme (IPS) claims payment. International Journal of Recent Technology and Engineering, 8(2 Special Issue 11), 660–663. https://doi.org/10.35940/ijrte.B1104.0982S1119
8. Chen, B., & others. (2026). From Prediction to Prevention: Using Text Mining and Explainable Machine Learning for Urban Bus Accident Analytics. Risk Analysis.
9. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
10. De Reyck, B., Degraeve, Z., & Vandenborre, R. (2008). Project options valuation with net present value and decision tree analysis. European Journal of Operational Research, 184(1), 341–355.

11. Gustman, A. L., & Steinmeier, T. L. (1999). Effects of pensions on savings: analysis with data from the health and retirement study. Carnegie-Rochester Conference Series on Public Policy, 50, 271–324. https://doi.org/10.1016/S0167-2231(99)00030-5
12. Jamal, N. F., Abdul Ghafar, N. M., Ismail, I. L., & Awang Chek, M. Z. (2018). Comparative study on the complex samples design features using SPSS Complex Samples, SAS Complex Samples and WesVarPc. International Journal of Academic Research in Business and Social Sciences, 8(4), 1282–1292. https://doi.org/10.6007/IJARBSS/v8-i4/4238
13. Jamal, N. F., Abdul Ghafar, N. M., Ismail, I. L., Awang Chek, M. Z., Ghafar, N. M. A., Ismail, I. L., & Chek, M. Z. A. (2018). Comparative study on the complex samples design features using SPSS complex samples, SAS complex samples and WesVarPc. International Journal of Academic Research in Business and Social Sciences, 8(4), 1282–1292. https://doi.org/10.6007/IJARBSS/v8-i4/4238
14. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2023). An Introduction to Statistical Learning with Applications in Python. Springer.
15. Khairuddin, M. Z. F., Sankaranarayanan, S., Hasikin, K., Razak, N. A. A., & Omar, R. (2024). Contextualizing Injury Severity from Occupational Accident Reports Using an Optimized Deep Learning Prediction Model.
16. Kose, I., Gokturk, M., & Kilic, K. (2015). An interactive machine-learning-based electronic fraud and abuse detection system in healthcare insurance. Applied Soft Computing Journal. https://doi.org/10.1016/j.asoc.2015.07.018
17. Koster, O., Kosman, R., & Visser, J. (2021). A Checklist for Explainable AI in the Insurance Domain.
18. LaValley, M. P. (2008). Logistic Regression. Circulation, 117(18), 2395–2399. https://doi.org/10.1161/CIRCULATIONAHA.106.682658
19. Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems.
20. Lutfi Abas, A., MohdSaid, A. R., Aziz Mohammed, M. A., & Sathiakumar, N. (2013). Use of a national reporting system for Occupational Diseases among Non-Governmental Employees in Peninsular Malaysia , 2002 to 2006. American Journal of Industrial Medicine, 56(1), 65–76.
21. Peng, C. J., & So, T. H. (2002). Logistic Regression Analysis and Reporting: A Primer. Understanding Statistics, 1(1), 31–70. https://doi.org/10.1207/S15328031US0101_04
22. Sahak, N. H., Mani, K. K., Hamedon, T. R., Rampal, K. G., & Vythialingam, K. (2020). Workers Survived from Commuting Road Crashes: Who Are They? International Journal of Academic Research in Business and Social Sciences, 10(15). https://doi.org/10.6007/IJARBSS/v10-i15/8342
23. SAS. (n.d.). Data mining in the insurance industry.
24. Selamat, M. N., & Surienty, L. (2015). An Examination of Commuting Accident in Malaysia.
25. Su, X., & Bai, M. (2020). Stochastic Gradient Boosting Frequency-Severity Model of Insurance Claims. PLoS ONE, 15. https://consensus.app/papers/stochastic-gradient-boosting-frequencyseverity-model-of-su-bai/0b4724fc7cfe50338888af148136a0a2/?utm_source=chatgpt
26. Woodfield, T. J. (2010). Predicting workers' compensation insurance fraud using SAS Enterprise Miner 5.1 with SAS Text Miner. Society of Actuaries (SOA). www2.sas.com/proceedings/sugi30/071-30.pdf
27. Zuwairy, M. S., Harith, A. A., Nobuyaki, H., Naim, N. M., & Yon, R. (2020). Road Traffic Accidents: A Descriptive Study of Commuting Injury Among Healthcare Workers in Malaysia, 2014--2016. International Journal of Public Health and Clinical Sciences.