Structures 71 (2025) 108005

Contents lists available at ScienceDirect

Structures

journal homepage: www.elsevier.com/locate/structures

# A supervised machine learning approach for structural overload classification in railway bridges using weigh-in-motion data

N.T. Le [a,b], M. Keenan [a,c], A. Nguyen [a,*], S. Ghazvineh [d], Y. Yu [e], J. Li [f], A. Manalo [a]

[a] University of Southern Queensland, Australia
[b] Hanoi University of Civil Engineering, Vietnam
[c] KiwiRail, New Zealand
[d] Kharazmi University, Iran
[e] University of New South Wales, Australia
[f] University of Technology Sydney, Australia

ARTICLE INFO

*Keywords:*
Weigh-in-Motion
Railway Bridge
Structural Overload Assessment
Axle Load Combination
Supervised Machine Learning

ABSTRACT

Weigh-in-motion (WIM) data provides valuable information on vehicle axle load, enabling efficient and economical railway structural safety management programs. However, the current method for assessing structural overload on railway bridges using WIM data is time-consuming and often requires line closure while analyses are being conducted. This paper presents the development of a novel supervised machine learning (ML) approach that can be used as an assessment tool to expedite the decision-making process and minimise economic loss. Variables for model input are carefully considered by analysing real WIM data obtained from measurement sites in New Zealand. Various supervised ML classification models are evaluated for their capability in classifying axle load combinations (ALC) into "Normal", meaning safe to go, or "Overload", meaning that line closure is required for detailed inspection of the affected bridges. It is found that the model using Neural Network (NN) outperforms other candidates in this capacity and is therefore selected for detailed model development. An initial investigation using a small dataset derived from real WIM measurements demonstrates that the NN model can achieve impressive evaluation metrics such as F1-score of 99.2 %. Subsequently, a method for artificially generating synthetic ALC data is proposed to create extensive training datasets for comprehensive structural overload model development. It is demonstrated that with sufficient overload data in the training dataset, the model can achieve an exceptional performance, reaching an F1-score of 99.84 % or higher for a single overload level and 99.5 % to 99.86 % for multiple overload thresholds. The developed model can be integrated into the WIM post-processing systems, providing a real-time bridge overload assessment tool that facilitates more efficient and cost-effective railway structural safety management.

## 1. Introduction

Railway transport infrastructure is crucial for any society as they provide an efficient, high-capacity mode of transport over long distances, facilitating the movement of goods and people, which is vital for economic growth and connectivity. Bridges are a key component of the railway network that help to overcome geographical barriers such as rivers and valleys. As reported by its Transport Agency [1], New Zealand has more than 1600 rail bridges, spanning more than 60 kilometres. However, a significant number of these bridges are of considerable age, on average being close to 80 years old, which means that they are reaching the end of their designed lifespans. Similar situations are reported around the world [2,3]. As these critical structures approach the end of their service life, designing cost-effective strategies for accommodating emerging transportation needs and mitigating potential dangers are essential [4]. Although structural health monitoring techniques can provide valuable insight into the bridge's resistant capacity from time to time [5-7], it is equally important to control the vehicle live

*Abbreviations:* AI, Artificial Intelligent; ALC, Axle Load Combination; BA, Boosting Algorithms; RBSO, Rail Bridge Structural Overload; B-WIM, Bridge Weigh-In-Motion; DT, Decision Tree; FN, False Negative; FP, False Positive; KNN, K-Nearest Neighbour; LR, Logistic Regression; ML, Machine Learning; NN, Neural Networks; SVM, Support Vector Machines; TN, True Negative; TP, True Positive; WIM, Weigh-In Motion; WIM_ALC, Axle Load Combination extracted from WIM data.

* Corresponding author.
*E-mail address:* andy.nguyen@unisq.edu.au (A. Nguyen).

https://doi.org/10.1016/j.istruc.2024.108005
Received 18 March 2024; Received in revised form 10 October 2024; Accepted 6 December 2024
Available online 12 December 2024

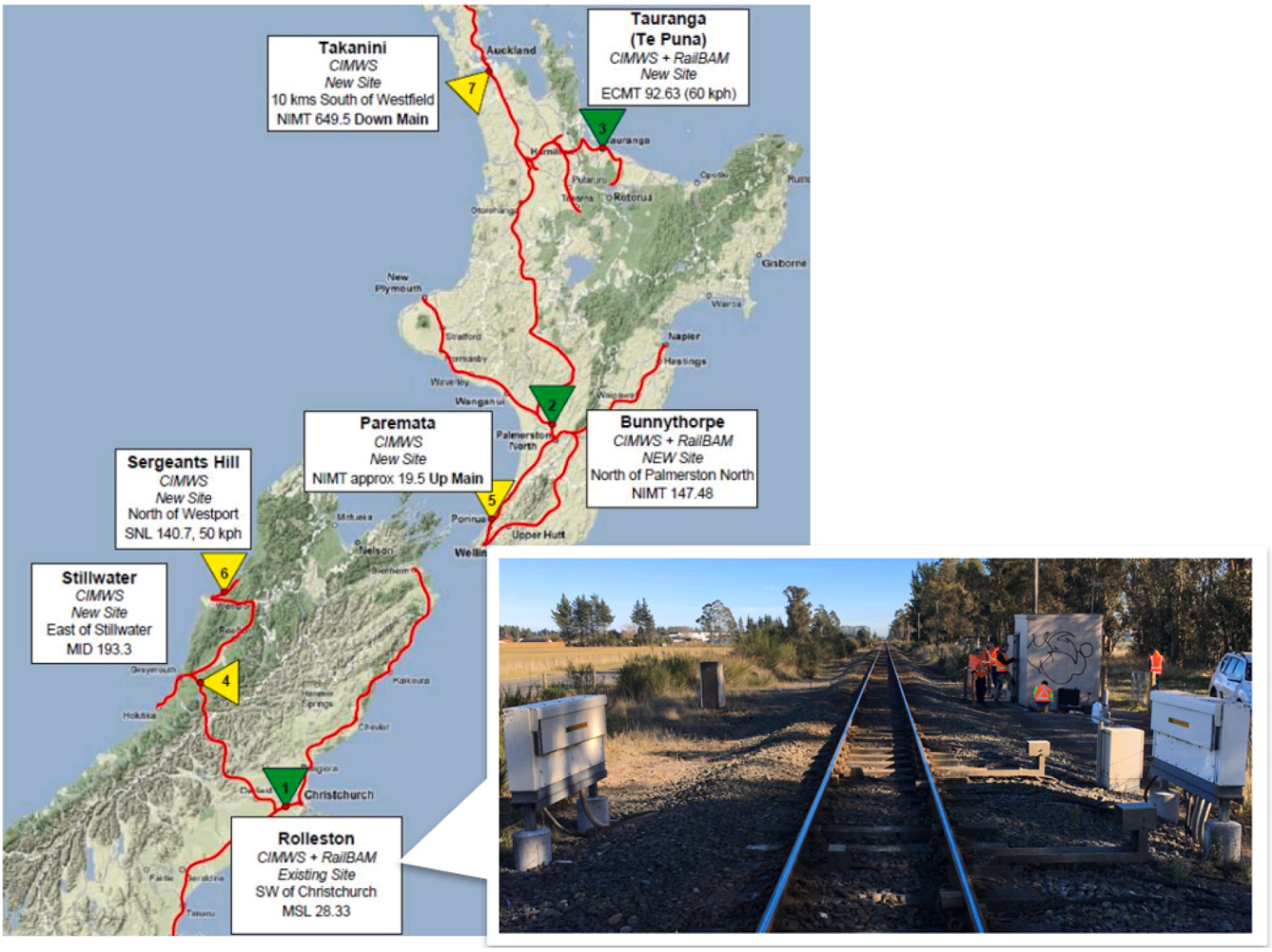


**Fig. 1.** Illustration of coupled in-motion weighing sites in New Zealand (Adapted from [27]).

loads to avoid structural overload since this has been nominated as one of the main causes of bridge failures [8]. Accordingly, monitoring the train axle loads is crucial to maintain the safety and longevity of the railway bridges.

Over the past decades, weigh-in-motion (WIM) has emerged as a viable alternative to the traditional static method for weighing transport vehicles. This shift can be attributed to the increasing demand for determining axle weights while the vehicles are in motion, especially in transport systems with high traffic volumes [9]. As a result, a large number of WIM technologies have been developed and put into use on a variety of transport networks, which can be classified into pavement-based WIM [10–13], bridge WIM (B-WIM) [14–21], and rail WIM [22–24] systems according to the type of structures on which they are installed. A typical WIM system includes sensors, data gathering devices, and data analysis algorithms that can compute the axle loads, axle spacing, gross vehicle weight, speed, and other factors of the moving vehicles [9]. Despite differences in the design, these WIM systems share a common purpose in providing valuable information on actual traffic loading for various structural health assessment and structural overload control purposes [17,21,23]. Recent advances in WIM technologies allow vehicle loads to be measured accurately, contributing to efficient and economical safety management of transport infrastructures. A properly installed and calibrated single pavement-based load cell WIM system can provide gross vehicle weights that are within 6 % of the actual vehicle weight for 95 % of the measured trucks [13,25]. Comprehensive rail weighing tests on a bridge WIM system in Poland reported accurate results, with errors within 5 % for 97 % of the test carriages, and within 2 % for 75 % of the test carriages [16]. Similarly, accurate measurement of train axle load can be achieved with an error margin of no more than 5 % and a mean absolute error of less than 2 % of the total vehicle weights using rail WIM systems [23].

The advances in WIM technology have enabled automation in measuring and controlling train axle loads on railway networks in many countries. Not long ago, New Zealand installed and upgraded its coupled in-motion weighing (CIMW) systems, a type of rail WIM as illustrated in Fig. 1, to incorporate sophisticated automatic vehicle identification and data processing algorithms, with the ability to provide instant alert on axle overload and imbalance conditions of train wagons [26]. The results are interfaced with the National Train Control Centre via the internet and have significantly reduced the risk of overloaded wagons in the country. In the event an individual wagon axle overload is identified by the WIM system, a structural overload assessment process will be activated in which structural engineers are tasked to examine the overload conditions on affected bridges (Fig. 2) [27]. However, this approach is rather time-consuming and most often, by the time structural assessment process finishes, the train controller has already made a number of operational actions such as stopping trains from continuing, contacting customers and rearranging freight movements, and coordinating structural inspections. All these can incur an extra cost of tens of thousands of dollars and other economic impacts as the result of the transport delay. Therefore, it is desirable to have a more efficient and timely structural overload assessment tool to minimise the duration of

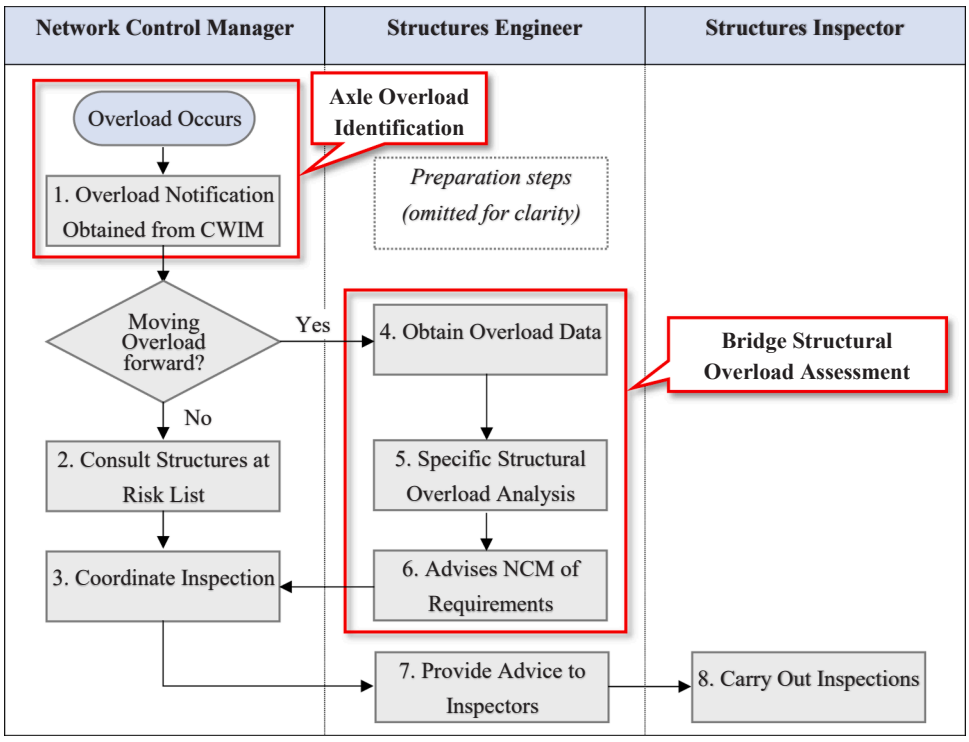


**Fig. 2.** KiwiRail bridge structural overload assessment process [27].

line closures and associated economic impact.

To address this need, an automated supervised machine learning approach for the assessment of overload condition in railway bridges under axle load combinations using measured data from nearby CIMW sites is developed. For demonstration of the concept, a standard 6 m bridge span length is chosen, and variables for model inputs are formulated from real WIM data obtained from measurement sites in New Zealand. Compared to the traditional rule-based approach, the machine learning approach offers greater flexibility especially when there is data variability due to operational and environmental factors. Another advantage of the machine learning approach is its ability to learn from long-term data to predict future conditions of the structure.

The content of the rest of the paper is as follows. Section 2 provides background on the current railway bridge structural overload assessment practice, along with an overview of machine learning (ML) classifiers and performance evaluation metrics employed in this study. Section 3 serves as a preliminary investigation into using supervised ML approach for structural overload assessment in railway bridges using WIM data. An original axle load combination dataset from real WIM raw data is created to characterise axle variables, select best-performing ML model, and to determine the optimal overload data proportion for maximizing classification accuracy. Section 4 presents a method that uses synthetic data to develop a robust supervised learning model capable of both binary and multi-class classifications. The paper then concludes with summary and final remarks on the study.

## 2. Background

### 2.1. Railway bridge structural overload assessment

The railway bridge structural overload (RBSO) assessment was conducted to evaluate whether external loads caused by passing trains exceed the bearing threshold for a specific bridge. The following subsections outline key aspects of current RBSO assessment practice in New Zealand as a case study, thereby highlighting potential areas for improvement that this research aims to address.

**Table 1**
KiwiRail's axle overload alert levels [28].

| Axle load range (V1 or V2) | Axle overload Alert levels | Action |
|---|---|---|
| 18t÷19.8t (Axle overload of 0 % to 10 %) | LOW | No action to be undertaken |
| 19.8t÷20.7t (Axle overload of 10 % to 15 %) | MEDIUM | Slow overload train to 40kph until unloaded. Stop other trains on the route until structural inspection completed. |
| 20.7t÷21.6t (Axle overload of 15 % to 20 %) | HIGH | Slow overload train to 25kph until unloaded. Stop other trains on the route until structural inspection completed. |
| > 21.6 t (Axle overload of greater than 20 %) | EXTREME | Slow overload train to 10kph until unloaded. Stop other trains on the route until structural inspection completed. |

#### 2.1.1. Axle overload and bridge structural overload

It is necessary to distinguish between axle overload on rail track and structural overload on rail bridges. Axle overload refers to the state when a specific axle load exceeds a designated threshold [12,28], which can potentially lead to damage in railway structures, including rail tracks, rail bridges, and other structural components. The focus of this paper is on structural overload on railway bridges, defined as the condition when a combination of axle loads traversing a bridge and causing additional bending moments and/or shear forces that exceed the bridge's live-load bearing limits determined through structural health monitoring programs [3,27,29]. The main difference between structural and axle overload lies in the consideration of axle load combinations (ALC) placed at different distances across a span length rather than the load under a single axle.

Practically, the allowable axle loads are determined on a line-by-line basis for each local rail network. For example, many train lines in New Zealand [28] currently have a maximum wagon axle load of 18 t, with a

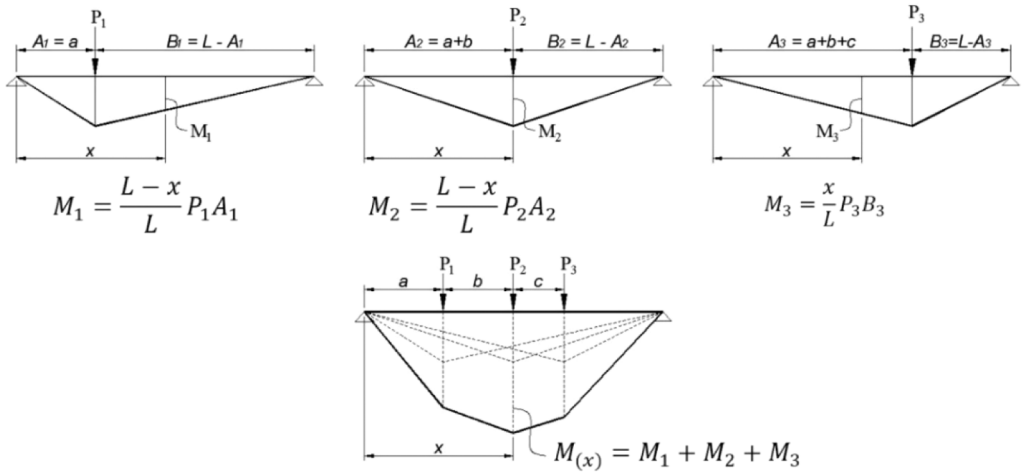


**Fig. 3.** Illustration of superposition-based bending moment calculation from three axle loads.

permitted excess of 10 % (19.8 t). Wagon axle weights exceeding 19.8 t trigger axle overload alerts, followed by various control actions depending on the level of exceedance, as summarised in Table 1. While WIM systems can immediately issue such axle overload alerts, engineers still need time to calculate the resulting forces on affected bridges and compare them to the live-load bearing limit to determine whether an RBSO has occurred [27]. Therefore, it is highly desirable to have an automated classification tool to assist this decision-making process with more rapid and accurate railway bridge structural overload warnings.

#### 2.1.2. Railway bridge structural overload assessment process

A number of coupled in motion weighing (CIMW) sites fitted with strain gauge-based sensors are installed throughout New Zealand (Fig. 1) to reduce the risk of overloaded wagons. Currently, KiwiRail, the country's railway network manager, evaluates structural overload for railway bridges utilising measurement data captured by the CIMW system in the form of axle forces and distances between them. The process following a wagon overload detection is shown in Fig. 2. In the case an axle overload occurs, immediate adoption of conservative protocol follows, leading to a line closure for 12 to 24 hours. As illustrated in Fig. 2, contributing to this line closure is the time required for the structural overload analysis (Step 4 to 6), inspection arrangement (Step 2, 3, 7), and inspection programs (Step 8) to be carried out in a number of at-risk bridges. Ideally, inspection arrangements (Step 3) should follow the confirmation of structural overload in the affected bridges (Step 6). However, under current conservative protocol, inspection preparation operations in step 3 and 7 force engineers and inspectors to approach at-risk bridge sites at the same time as the engineers are analysing the overload data. This is attributed to the substantial time lag in the current analytical overload assessment process. Developing of a more robust RBSO assessment tool to evaluate overload condition immediately after Step 1 could help eliminate unnecessary inspection programs, thereby reducing temporary line closure duration.

#### 2.1.3. Analytical method for RBSO assessment

In the current RBSO assessment process, KiwiRail employs an analytical approach for evaluating structural overload condition in railway bridges (Fig. 2, Step 5). The backbone of this approach is an algorithm to calculate the maximum bending moment and shear forces of bridge girders under moving ALCs [30]. For the sake of completeness, Fig. 3 illustrates the principle of the superposition method used by KiwiRail to calculate the bending moment of bridge girders with span length $L$ under a combination of three axle loads. In this example, individual moments at point $x$ are calculated for each load before being summed to establish the total moment from all loads. The calculation of shear forces follows an analogous procedure, which is not presented in this paper for brevity.

The challenge lies in identifying the actual position of the ALC and the position of girder's section that returns the highest overall moment, particularly given the variation in both axle weights and spacings along the train being considered. The shorter the bridge span length, the easier and faster the calculation process is because the number of axles on it decreases, and vice versa. The best way to address this challenge is to effectively iterate the process, i.e., move the ALC positions incrementally across the span keeping the axles at the same consistent separation, recording the total moments progressively and selecting the maximum moment ($M_{\max}$) after the process. It should be noted that for each ALC move, another iteration is carried out at different girder sections to find the position with maximum moment value of that move. Realistically, as the maximum moment is always at or near the middle of the span ($x = L/2$), the moments only need to be calculated over the middle quarter of the span [3,30].

It is evident that, to reach highly accurate $M_{\max}$ results, the above double-iteration process should be carried out at very small intervals for each of the ALC moves and for each of the maximum moment searches. This is undoubtedly a time-consuming process given the large number of affected bridges and the number of ALCs to be assessed. One way to circumvent this issue is through programming. However, for complex bridges such as multiple span and truss bridges, this solution is impractical since closed-form formulae of the internal forces do not exist.

In summary, the superposition method serves as a fundamental element in KiwiRail's current structural overload assessment of railway bridges. Although this method is based on solid mechanical principles, its limitations highlight the need for a more rapid automated approach integrated into the WIM post-processing system, which the present paper aims to address. In addition, this study focuses solely on bridge assessment using bending moment capacity, since the bending stress is commonly the governing stress in the bridge girders.

### 2.2. Supervised machine learning-based classification models

#### 2.2.1. Machine learning for structural overload assessment

Machine learning (ML), a branch of artificial intelligence (AI), leverages sophisticated algorithms to extract knowledge from data. This process enables the system to make informed predictions or decisions without being explicitly programmed. Instead, the system is trained using data, allowing it to independently adapt and improve its performance over time. A comprehensive review of state-of-the-art ML algorithms for structural engineering can be found in [31]. ML can be an

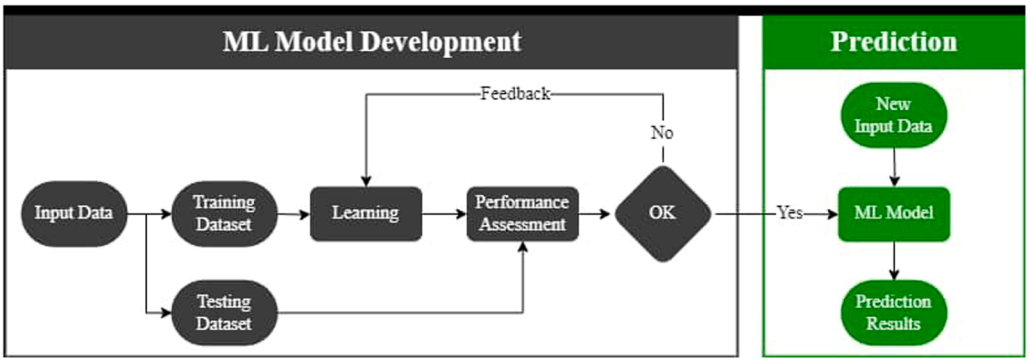


**Fig. 4.** Workflow of Machine Learning (ML).

effective tool for automated assessment of structural overload in railway bridges, thereby addressing the shortcomings in the current analytical RBSO assessment method as identified in the previous section.

The computer system (i.e., the ML model) undergoes a learning process to make accurate predictions when presented with new data [32]. Fig. 4 illustrates a typical workflow for ML in predictive modelling. The process begins with an initial dataset and a learning algorithm, through which the computer system is trained and validated to improve its performance iteratively until it achieves the desired level of accuracy. Once the model passes the development process, it is then deployed for making predictions with new data.

ML algorithms can be classified into several categories based on the learning process. The most fundamental type of ML is supervised learning, whereby algorithms are trained using a labelled dataset to predict outcomes and recognize patterns by learning the relationship between the inputs and the outputs. An unsupervised learning algorithm, on the other hand, discovers patterns in unlabelled data without any explicit guidance or instruction. Another class of ML is semi-supervised learning, which is similar to supervised learning but it uses both labelled and unlabelled data. Using this combination, ML algorithms can learn to label unlabelled data. This approach is suitable when only a small portion of the data is labelled and determining true labels for the rest of the data is expensive.

With 'normal' and 'overload' as the binary labels, the assessment exercise herein is apparently a classification problem, as opposed to a regression problem, in supervised learning. The next subsection will provide background on the classification models commonly used in supervised learning.

### 2.2.2. *Supervised classification models*

In supervised learning, the process begins with training the model using a labelled dataset known as the training data. During this training phase, the model learns to understand the patterns and relationships between the input features and their corresponding labels. Once the model has been trained, it is then tested using a separate dataset, known as the testing data. This data is used to evaluate the performance of the model and its ability to generalize to unseen data. After the model has been trained and tested, it can then be used to predict the categorical label of new, unlabelled data. In essence, the model applies the understanding it gained during the training phase to classify the new data based on its features. Various supervised classification algorithms are available for this purpose, such as Logistic Regression Analysis (LR), K-Nearest Neighbour (KNN), Support Vector Machine (SVM), Decision Trees (DT), Boosting Algorithms (BA), and Neural Networks (NN) [31]. The following is a brief description of the algorithms with advantages and disadvantages highlighted. Interested readers can refer to [31] for more references.

LR is a simple and widely used technique for binary classification problems based on probabilistic analyses. The model applies a sigmoid function to the output of a linear model and estimates the probability of an instance belonging to a certain class. It is easy to implement and interpret, but it may suffer from underfitting and multicollinearity issues.

KNN and SVM are two popular distance-based classification methods in ML. KNN classifies data points based on the class of their K-nearest neighbours in the training data. It is simple and intuitive, but sensitive to noisy data and curse of dimensionality. SVM finds the optimal hyperplane (or decision plane) to separate two classes in high-dimensional spaces. The technique is robust against outliers and effective for complex data but may be computationally expensive and sensitive to hyperparameter tuning.

DT segment data into smaller subsets based on specific criteria. They can handle both categorical and numerical features, as well as manage missing values. However, they are prone to overfitting, i.e., memorising the training data and losing the ability to generalise with new data.

BA enhance weak learners by sequentially adding them to the model. Examples such as AdaBoost and Gradient Boosting are effective for complex tasks but can be computationally expensive. Extreme Gradient Boosting is a popular and efficient method that uses gradient boosting to optimise decision trees. It can handle large-scale and sparse data, as well as imbalanced classes. Although it has many parameters that can be tuned to improve the performance, it may also be prone to overfitting and thus requires careful validation.

Finally, NN is among the most complex and powerful methods that can learn nonlinear and high-level features from the data. NN can handle various types of data, such as images, text, or audio. Leveraging logistic regression within their classification layers, NN enhances their accuracy in modelling binary outcomes. They typically require a large amount of data and computational power, as well as careful design and training. Due to their complex structure, they may also suffer from overfitting, underfitting, or being stuck in local minima.

All the above supervised classification model algorithms have their own advantages and limitations and are promising for application to the ALC classification problem. This research will examine the performance of these ML models to find the most appropriate algorithm for automated RBSO assessment using WIM data.

### 2.2.3. *Performance evaluation metrics*

In this study, the performance of the classification models is evaluated through five evaluation metrics: accuracy, precision, recall, F1-score, and the area under the curve of the receiver operator characteristic (AUC-ROC). Accuracy is the most popular metric for ML model evaluation and selection, which measures how often the classifier correctly predicts the unseen data. It can be defined as the ratio of the number of correct predictions to the total number of predictions, which is formulated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \times 100\% \quad (1)$$

where, true positive (TP) is the number of samples being correctly

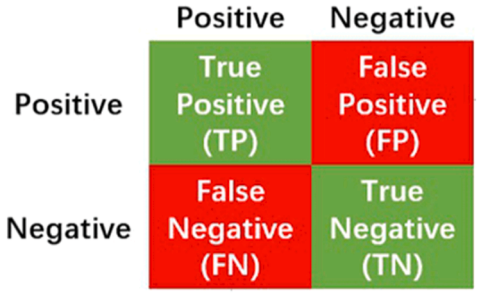


**Fig. 5.** Confusion matrix.

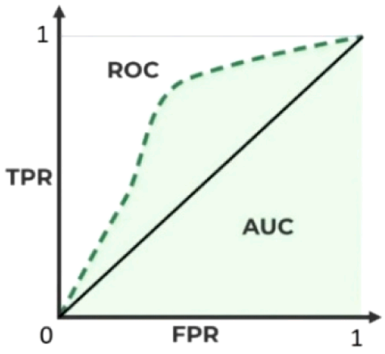


**Fig. 6.** ROC-AUC classification evaluation metric.

classified as positive, false positive (FP) is the number of positive samples being misclassified in the negative class, false negative (FN) shows the number of positive samples incorrectly categorized as negative, and true negative (TN) provides the number negative samples correctly classified. In binary classification problems, these indices can be expressed in the form of confusion matrix (Fig. 5), which is a convenient way to visualise the classification results.

Accuracy is useful when the target class is well balanced. For imbalanced data, other metrics should be considered. Precision is a useful matrix when a high FP rate is a concern. It is defined as the fraction of the correctly identified positive instances among all the positive instances retrieved (Eq. (2)). When false negative control is important, the recall index is useful. As formulated in Eq. (3), recall is the fraction of the TP prediction instances among the total positive instances of the sampling data. It is often convenient to combine precision and recall into a single metric called F1-score, which is the harmonic mean of the two metrics. F1-score ranges from 0 (unable to classify) to 1 (perfect classification) as can be inferred from Eq. (4). Compared to accuracy, F1-score is more sensitive to false detection and is more suitable for detailed model evaluation.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F1} \quad = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

The AUC-ROC is another useful metric for model selection when the data is imbalanced. Receiver operator characteristic (ROC) is a probability curve that plots the true positive rate (TPR=TP/(TP+FN)) against the false positive rate (FPR=FP/(TN+FP)). The area under the curve of the ROC is a useful representation of binary classification model performance at various classification thresholds (Fig. 6). It is used to assess the ability of a classifier to distinguish between classes (e.g., normal or overload). The area always falls within 0 and 1, and a greater value of AUC denotes better model performance.

## 3. Initial model development for RBSO assessment

This section serves as a preliminary investigation on using supervised ML approach for structural overload assessment in railway bridges using WIM data. A weigh-in-motion axle load combination (WIM_ALC) dataset was created from real WIM data and used to select the most suitable ML model for RBSO classification. The selected ML model's performance was then evaluated on untrained data and its sensitivity to varying proportions of overload data. The findings of this investigation will inform the final model development in Section 4.

### *3.1. Creation of a WIM_ALC dataset for structural overload assessment*

#### *3.1.1. The raw WIM axle load data*

The original WIM measurement data used in this study was obtained from KiwiRail. A dataset equivalent to three months of operation was collected from a typical WIM site like the one depicted in Fig. 1. Permission to use this authentic data has been granted, and a large amount of historical data is available in tabular format, as shown in

**Table 2**
Samples of the raw WIM data (to be viewed with Fig. 7).

| Train Time | Vehicle Tag | Train Axle | Axle No. | Axle Pitch (mm) | Axle Weight (t) | Bogie Avg. Axle Weight (t) |
|---|---|---|---|---|---|---|
| **20xx-xx-xx 15:27:29** | CE 003195 | 33 | 1 | 2060 | 16.71 | 16.4 |
| **20xx-xx-xx 15:27:29** | CE 003195 | 34 | 2 | 1770 | 16.13 | 16.4 |
| **20xx-xx-xx 15:27:29** | CE 003195 | 35 | **3** | 8600 | **17.02** | **17.2** |
| **20xx-xx-xx 15:27:29** | CE 003195 | 36 | **4** | **1750** | **17.29** | **17.2** |
| **20xx-xx-xx 15:27:29** | CE 000198 | 37 | **1** | **2080** | **16.21** | **15.9** |
| **20xx-xx-xx 15:27:29** | CE 000198 | 38 | **2** | **1770** | **15.55** | **15.9** |
| **20xx-xx-xx 15:27:29** | CE 000198 | 39 | 3 | 8630 | 16.91 | 17.1 |
| 20xx-**xx**-xx 15:27:29 | CE 000198 | 40 | 4 | 1750 | 17.19 | 17.1 |

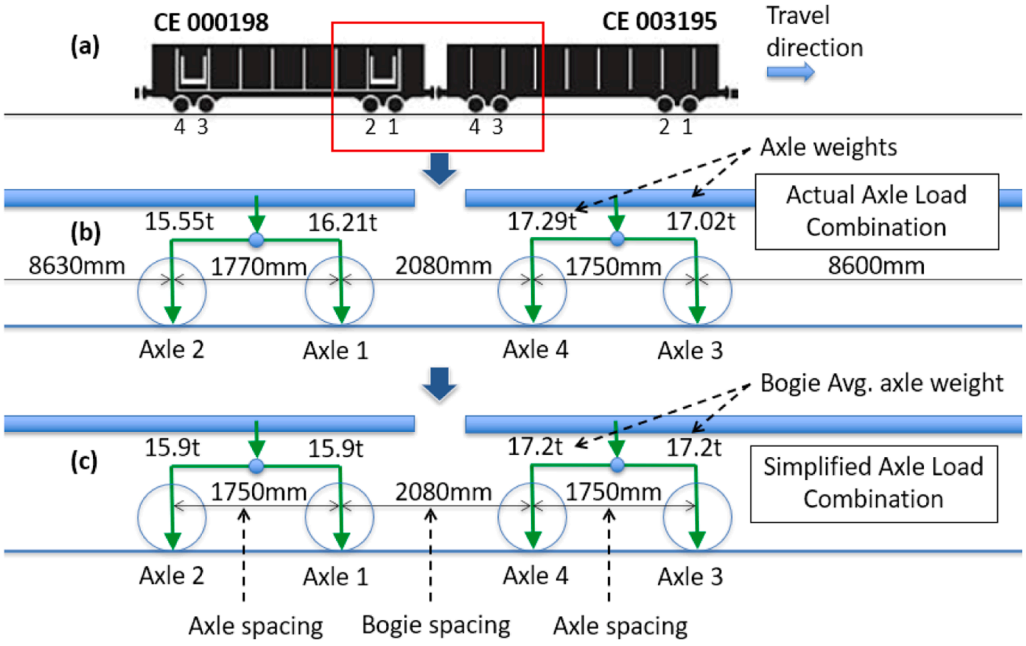


**Fig. 7.** Illustration of an Axle Load Combination.

Table 2.

For brevity, only information related to wagon axle loading is shown in Table 2. The data rows are arranged in reverse chronological order, meaning that the last recorded axle is on the top row of the table. The first column of the WIM data indicates the date and time of the passing trains, followed by the "vehicle tag" column showing labels of the vehicles (wagons and locomotives). The first two or three letters of the vehicle tag designates the wagon type. The next column "Train Axle" displays the train's axle numbers, which starts from 1 for every train and increases up to the total number of axles of the train. The next column named "Axle No." depicts the axle numbers for each wagon in the order of 1–2-3–4 or 4–3-2–1 depending on the travel direction (Fig. 7-a). The last three columns of Table 2 store important axle load information regarding the axle pitch (spacing), axle weight, and the bogie average axle weight.

Fig. 7 visualizes the WIM data from Table 2. General information regarding vehicle tags, axle numbers and travel direction of the two adjacent wagons are shown in Fig. 7-a. Details of an axle load combination (ALC) constituted from the adjacent bogies of the two wagons are depicted in Fig. 7-b, with values taken from Table 2 highlighted in dashed frames. This is an example of ALC for structural overload assessment of railway bridges. In this study, a simplified ALC scheme is used where the average bogie axle weights and the smaller axle spacing are taken, which helps to reduce the variable number in the structural analysis (Fig. 7-c). The bogie axle weight was reasonably rounded down to one decimal place, which primarily serves to minimise model training costs. In addition, taking the average axle weights has been in common practice since measurement error on a total bogie load was proven to be smaller than the error on individual axles because the effects of imbalance are compensated for [22,24]. Moreover, using the smaller axle spacing is acceptable since the difference between the two axle spacing was found to be minimal in the recorded dataset.

### 3.1.2. *Axle load combination extraction and characterization*

A simple algorithm was created and used to extract the wagon axle load combinations from the raw WIM axle load data described in the previous section to create a training dataset, namely WIM_ALC, for rail bridge structural overload assessment in this study. The ALCs were extracted separately from one train to another, by scanning the train's axle number, which always starts from "1" for a new train (3rd column,

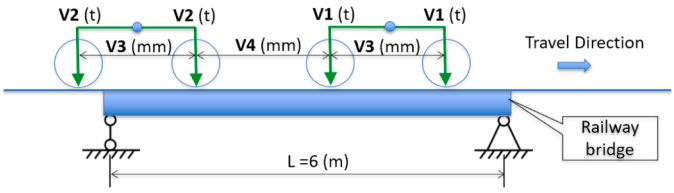


**Fig. 8.** Generalized 4-Variable ALC for bridge overload assessment.

**Table 3**
Examples of WIM_ALC data extraction.

| Train Time | Wagon Pair | Train Axle | V1 (t) | V2 (t) | V3 (mm) | V4 (mm) | $M_{max}$ (kNm) |
|---|---|---|---|---|---|---|---|
| 20xx-xx-xx 15:27:29 | CE-CE | 35 –38 | 17.2 | 15.9 | 1750 | 2080 | 430.38 |
| 20xx-xx-xx 19:55:56 | CE-CE | 59 –62 | 18.9 | 17.8 | 1790 | 2020 | 475.88 |
| 20xx-xx-xx 19:55:56 | CE-CE | 63 –66 | 18.0 | 18.1 | 1800 | 2030 | 458.64 |
| 20xx-xx-xx 20:56:13 | US-PK | 51 –54 | 9.9 | 12.5 | 1660 | 2570 | 287.23 |
| 20xx-xx-xx 21:24:27 | IH-IH | 23 –26 | 10.0 | 9.7 | 1730 | 2740 | 223.00 |
| 20xx-xx-xx 21:24:27 | PK-PK | 71 –74 | 12.4 | 12.7 | 1640 | 2820 | 284.71 |
| 20xx-xx-xx 16:53:11 | CB-CB | 19 –22 | 4.0 | 3.9 | 1740 | 2050 | 101.79 |
| 20xx-xx-xx 19:52:56 | IA-IH | 23 –26 | 9.7 | 5.1 | 1710 | 2770 | 209.97 |
| 20xx-xx-xx 8:44:30 | CE-CE | 79 –82 | 18.3 | 17.7 | 1760 | 2010 | 466.62 |
| 20xx-xx-xx 8:44:30 | CE-CE | 87 –90 | 17.3 | 17.7 | 1750 | 1980 | 455.62 |

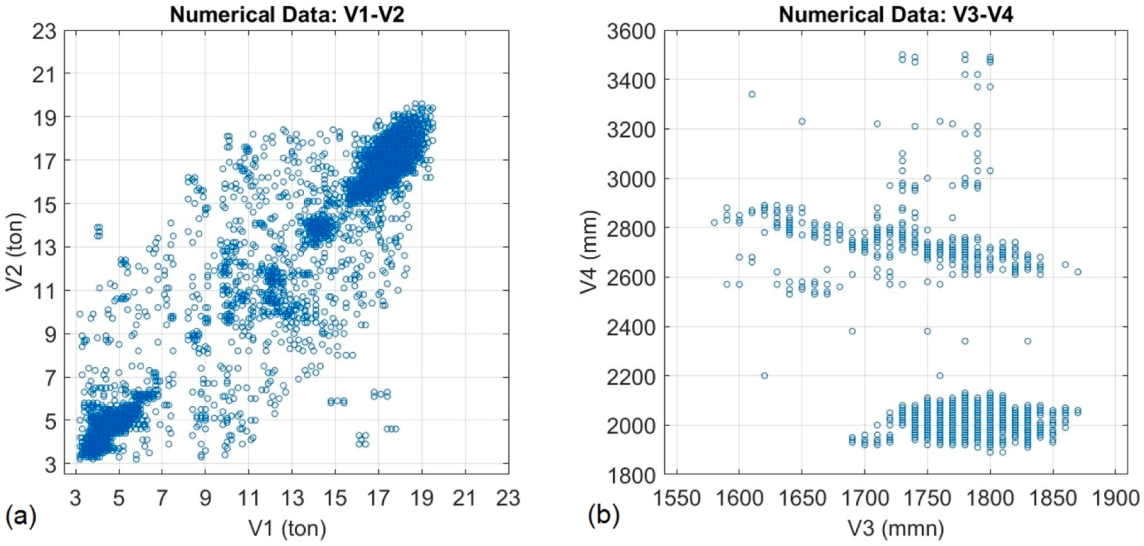


**Fig. 9.** Pair distribution of the ALC variables.

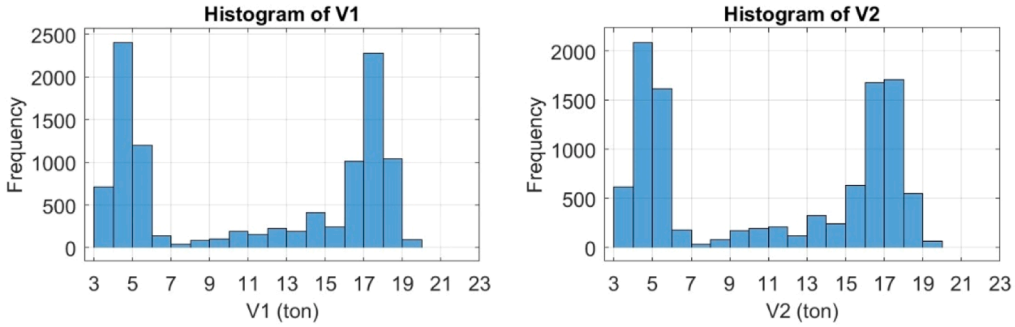


**Fig. 10.** Histogram of bogie axle loads (V1 and V2).

Table 2). In addition, only freight wagons were included in the training dataset; this was done by scanning the vehicle tags (2nd column, Table 2), to exclude locomotive axles.

After all wagon axles of one train were identified, the algorithm created all ALCs following a procedure as visualized in Fig. 7. Since this article focuses on 6m-long single-span railway bridges (which is among the most common rail bridges span in New Zealand) for proof of concept without losing generality, based on engineering judgement, a condition was set to consider only ALCs consisting of the last 2 axles and the first 2 axles of the two adjacent wagons, as depicted in Fig. 7-c. The ALC was then generalized into a 4-Variable ALC scheme as shown in Fig. 8, including two pairs of axle loads (V1, V2), a pair of axle spacing (V3), and a bogie spacing (V4).

The resultant ALC data, after removing duplicate combinations, contained over ten thousand rows of ALC samples representing various wagon types as illustrated in Table 3. The first three columns of the table store information of the train time, wagon types, and train axles for which the load combinations were created. The next four columns are the axle loads and spacings corresponding to the four ALC variables, as structured in Fig. 8. The first row of Table 3 is an example of the ALC extracted from Table 2 for the axles No.35 to No.38 of the two CE wagons crossing the WIM station on 20xx-xx-xx at 15:27:29 PM (xx denotes the year, month and day information that is withheld for confidentiality).

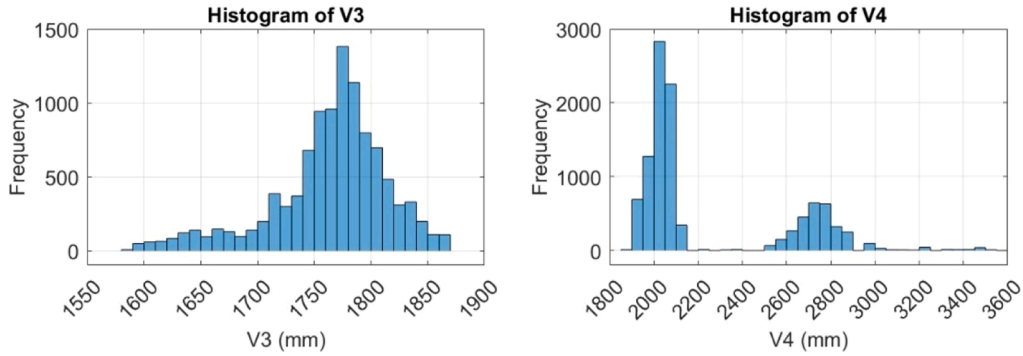


**Fig. 11.** Histogram of bogie axles spacing (V3) and wagon axle spacing (V4).

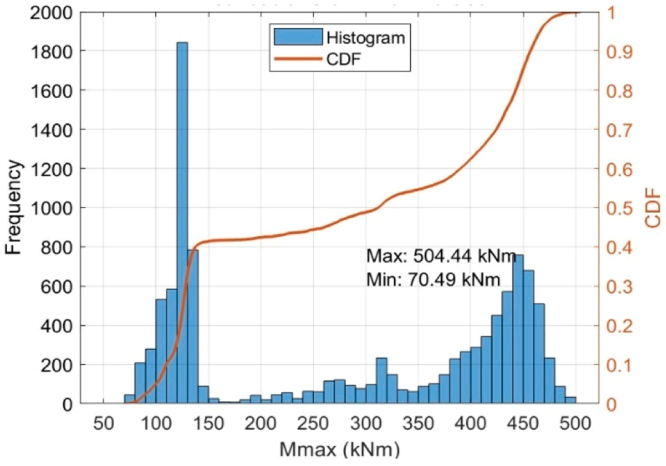


**Fig. 12.** Distributions of $M_{max}$ values of the WIM_ALC dataset.

Next, an iteration algorithm was created to calculate the maximum bending moment ($M_{max}$) induced by the ALCs for the selected 6-m-long span of simply supported bridge girders following the superposition method presented in Section 2.1.3. The resultant $M_{max}$ values were then added to the last column of Table 3. To give an insight into the WIM_ALC variable characteristics, the relationship between axle variables is extracted and shown in Fig. 9, and the distributions of these variables are shown in Fig. 10 and Fig. 11.

As shown in Fig. 9-a and Fig. 10, the collected data contains wagon axle loads ranging from 3 tons to 19.5 tons. V1-V2 pairs located near the diagonal line in Fig. 9-a represent axles under two adjacent freight wagons with approximately equal loads, while V1-V2 pairs located off the diagonal line indicate axles under wagons loaded differently. Similarly, the relationship between the axle spacing (V3) and the bogie spacing (V4) is shown in Fig. 9-b, and the distributions of these variables are presented in Fig. 11. It is evident that the bogie axle spacing V3 varied in a small range of [1550÷1900] mm, with a higher concentration around 1770 mm. By contrast, the bogie spacing V4 exhibited a wider range of values, from 1900 mm to 3600 mm. Higher concentrations of V4 were observed in the ranges of [1900÷2100] mm and [2500÷2900] mm. These characteristics are essential for the generation of synthetic ALC data later in Section 4.

To visualize the frequency distribution of the $M_{max}$ values across the current dataset, a histogram chart is plotted in Fig. 12. It can be seen that $M_{max}$ varies from 70.49 kNm to 504.44 kNm, with highest frequency in the ranges of 100÷130 kNm and 400÷470 kNm, corresponding to the empty and full-loaded states of the freight wagons, respectively. In addition, a cumulative distribution function (CDF) of $M_{max}$ is plotted to the right vertical axis of Fig. 12 to envision the variable probability distribution, which will be used later in Sections 3.1.3 and 3.2.3 to extract the $M_{allow\text{-}live}$ thresholds under different ALC overload portions.

### 3.1.3. Data labelling for preliminary model development

As stated in Section 2.1.3, the ALCs were classified in this study based on the additional bending moment they cause to the bridges as the main live load without the need to consider existing moment caused by other load types. Based on the calculated $M_{max}$ value, each ALC will be classified as "normal" if its $M_{max}$ is below an allowable bending moment value ($M_{allow\text{-}live}$) considering only live load predetermined for each bridge. Otherwise, the ALC will be classified as "overload". The $M_{allow\text{-}live}$ of an existing railway bridge depends on its traffic load rating condition, which shall be specified by the local rail network authority and may be subject to changes from time to time. Since the collected WIM data contains only axle loads up to 19.5 t, which fell into the allowable operational range (Table 1), the current WIM_ALC dataset should contain only "normal" class. To evaluate the feasibility of ML approach for assessing RBSO using WIM data, in this section, we prioritized dataset diversity over strict adherence to real traffic load rating conditions. Accordingly, a 5 % portion of the WIM_ALC data with the highest $M_{max}$ values was classified as overload, while the remaining 95 % was considered normal. Based on the CDF curve in Fig. 12, the corresponding $M_{allow\text{-}live}$ threshold of 465.47 kNm was calculated. Therefore, ALCs in Table 3 were categorized as 'normal' and labelled as '0' if their maximum bending moment $M_{max}$ was less than or equal to 465.47 kNm. Conversely, ALCs were classified as 'overload' and labelled as '1'. The WIM_ALC dataset was then restructured for ML model training, as shown in Table 4. The model predictors for training and testing were the four-axle variables (V1, V2, V3, V4), while the model response was the labelled structural overload status. The $M_{max}$ values in the last column were kept for reference only and not used for training or testing. The WIM_ALC dataset now contains 10,538 ALCs samples, of which 10,011 samples belong to the "normal" class, and the remaining 527 cases (5 %) belong to the "overload" class.

**Table 4**
Example of the WIM_ALC dataset for training and testing.

| V1 | V2 | V3 | V4 | True_Class | $M_{max}$ |
|---|---|---|---|---|---|
| 16.8 | 16.7 | 1810 | 2040 | 0 | 424.02 |
| 17.1 | 17.1 | 1800 | 2030 | 0 | 433.76 |
| 17.9 | 17.1 | 1810 | 2040 | 0 | 448.45 |
| 17.2 | 16.0 | 1810 | 2020 | 0 | 430.42 |
| 17.1 | 17.2 | 1810 | 2010 | 0 | 436.61 |
| 17.6 | 17.1 | 1800 | 2020 | 0 | 444.84 |
| 18.2 | 17.3 | 1770 | 2000 | 0 | 462.52 |
| 18.5 | 18.1 | 1750 | 2010 | 1 | 473.67 |
| 11.9 | 11.5 | 1820 | 2710 | 0 | 261.39 |
| 5.7 | 5.5 | 1760 | 2720 | 0 | 126.69 |

**Table 5**
Model validation information.

| Parameter | Value | Note |
|---|---|---|
| Observation | 10,538 | The data size |
| Number of Predictors | 4 | V1, V2, V3, V4 |
| Response Classes | 2 | [0,1] |
| Validation | k-fold, k = 5 | Cross Validation technique |
| Number of models | 33 | Models that can be quickly implemented in MATLAB Classification Learner App |

**Table 6**
Performance of the six best classification models.

| Model No. | Model type | Validation | | Training time (sec.)* | Model size (bytes) |
|---|---|---|---|---|---|
| | | Accuracy % | Total Cost | | |
| 1 | Fine Decision Tree (DT) | 98.94 % | 112 | 9 | 18,819 |
| 2 | Binary GLM Logistic Regression (LR) | 99.06 % | 99 | 13 | 1,305,242 |
| 3 | Quadratic SVM (SVM) | 99.26 % | 78 | 353 | 13,652 |
| 4 | Fine KNN (KNN) | 98.92 % | 114 | 10 | 780,558 |
| 5 | Boosted Trees Algorithm (BA) | 99.37 % | 66 | 14 | 262,320 |
| **6** | **Wide Neural Network (NN)** | **99.90 %** | **11** | **122** | **10,815** |

Note:
* Training performed by a personal computer

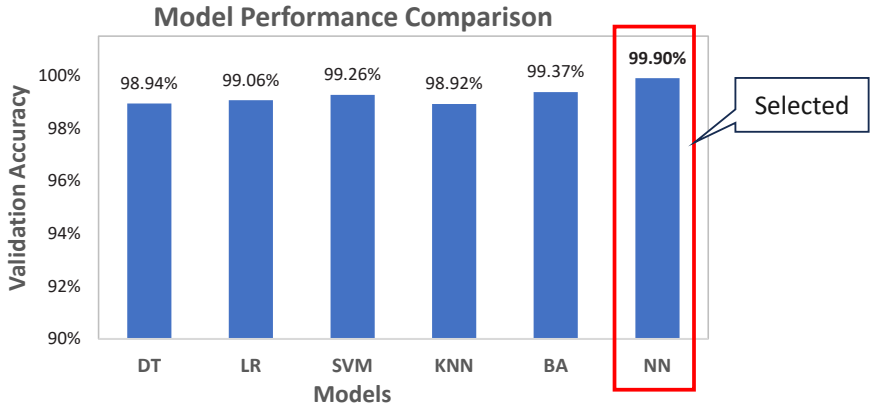


**Fig. 13.** Model performance comparison (to be seen with Table 6).

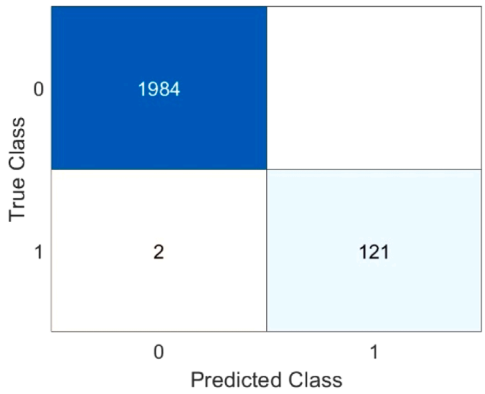


**Fig. 14.** A typical confusion matrix of the classification result.

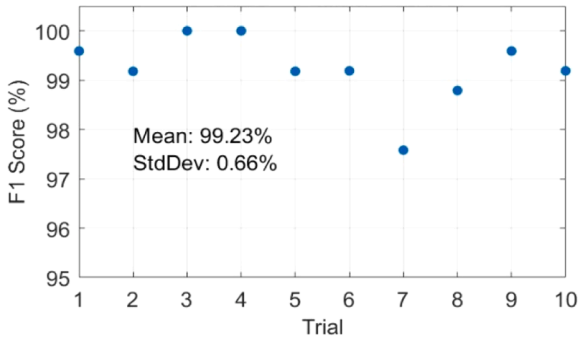


**Fig. 15.** A typical F1-score distribution among ten trials of one experiment.

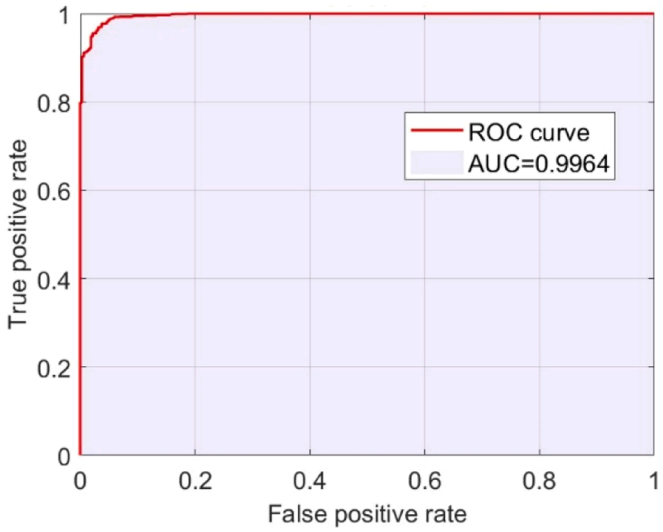


**Fig. 16.** Example of AUC-ROC result.

### *3.2. Development of RBSO Assessment Model Using the WIM_ALC Dataset*

#### *3.2.1. Selection of ML Model for RBSO Assessment*

The first step in the model development process for axle load combination classification was to evaluate the most commonly used machine learning models belonging to six main supervised ML classification groups described in Section 2.2.2. This was quickly deployed in the MATLAB Classification Learner App and the K-fold cross-validation technique was employed for maintaining the robustness of the model validation process. The WIM_ALC dataset was fed into thirty-three models that are available in the App. The input information for this comparison process is summarised in Table 5. After completing the training and validation process, the six best models representing each classification group were selected for comparison in terms of validation accuracy rate, total cost (total number of false detections), training time, and model size (Table 6). A plot of performance accuracy index among the six models is shown in Fig. 13. The results reveal that the best-performing model was the wide Neural Network (NN), with the highest classification accuracy of 99.90 %, lowest total cost, reasonable training time and model size. The NN model was therefore selected for model development for RBSO assessment.

#### *3.2.2. Performance of the neural network model within the WIM_ALC dataset*

After the NN classification model was selected, it was examined further to evaluate its ability in classifying untrained data within the WIM_ALC dataset using the hold-out cross-validation technique. It is a common practice to hold out part of the data as a test set to avoid overfitting when performing supervised machine learning experiments. The WIM data was therefore randomly partitioned into a training subset containing 80 % of the total samples and a subset of the remaining 20 % held for testing. This partitioning operation was repeated five times to form five separate experiments. In each experiment, the training-testing process was carried out 10 times to exclude outliers and to derive average performance indices. A typical testing result in the form of a confusion matrix is displayed in Fig. 14, with predicted results TN= 1984, TP= 121, FN= 2, FP= 0. Based on the predicted outcomes, the four confusion-based performance indices were calculated following Eqs. (1) to (4). For illustration, the F1-Score values of Experiment 5 are

**Table 7**
Cross validation results of NN model among the five experiments.

| Performance Index | Test 1 | Test 2 | Test 3 | Test 4 | Test 5 | Average |
|---|---|---|---|---|---|---|
| Accuracy | 99.93 % | 99.92 % | 99.91 % | 99.91 % | 99.91 % | 99.92 % |
| F1-Score | 99.37 % | 99.21 % | 99.13 % | 99.19 % | 99.23 % | 99.23 % |
| Precision | 99.92 % | 99.17 % | 99.78 % | 98.87 % | 99.04 % | 99.36 % |
| Recall | 98.83 % | 99.27 % | 98.49 % | 99.52 % | 99.43 % | 99.11 % |
| AUC | 99.77 % | 99.79 % | 99.73 % | 99.76 % | 99.99 % | 99.81 % |

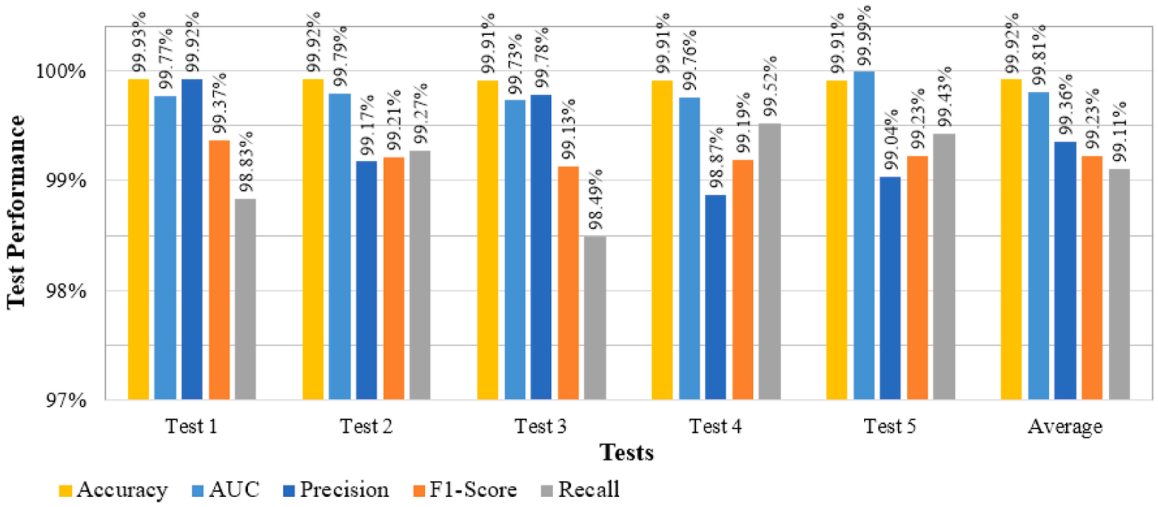


**Fig. 17.** Performance of the NN model using WIM_ALC data.

**Table 8**
Portions of dataset with different $M_{max}$ thresholds.

| WIM_ALC datasets | **Original** | **1** | **2** | **3** | **4** | **5** | **6** |
|---|---|---|---|---|---|---|---|
| Overload proportion (%) | 5 % | 10 % | 15 % | 20 % | 30 % | 40 % | 50 % |
| Normal load proportion (%) | 95 % | 90 % | 85 % | 80 % | 70 % | 60 % | 50 % |
| Equivalent $M_{allow\text{-}live}$ (kNm) | 465.47 | 456.79 | 449.42 | 442.83 | 424.30 | 389.95 | 311.40 |

plotted in Fig. 15, with an impressive average value of 99.23 % and a low standard deviation of 0.66 %. In addition, the ROC curve was extracted from testing results, from which the area under curve AUC metric was calculated and shown in Fig. 16. Average values of the five performance indices are summarised in Table 7 and plotted in Fig. 17.

For comparison purposes, all values are presented as percentages. The Accuracy and AUC indices, averaging 99.92 % and 99.81 %, respectively, demonstrated consistent performance and outperformed the other three metrics. While Precision and Recall were slightly lower at 99.36 % and 99.11 %, the F1-Score of 99.23 % indicated satisfactory

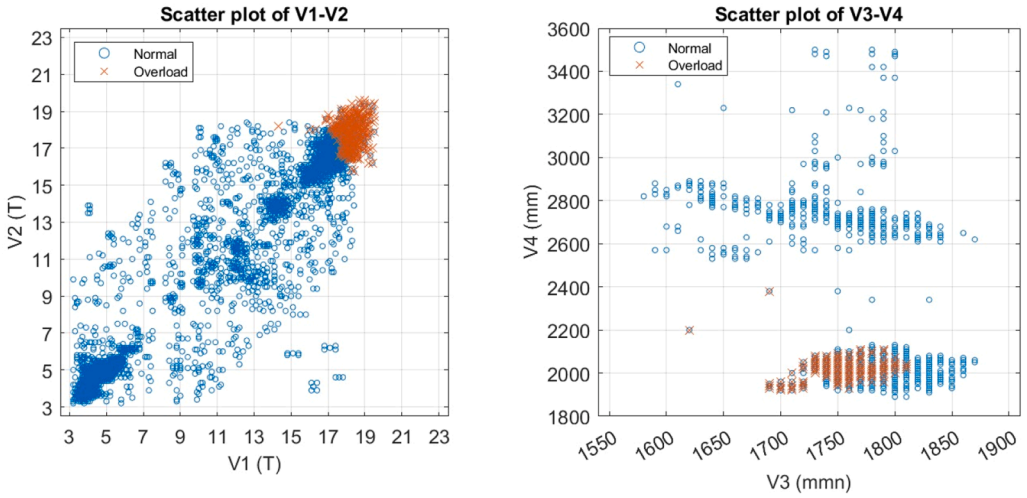


**Fig. 18.** Variable distribution of subset WIM_ALC_1 (OP=10 %).

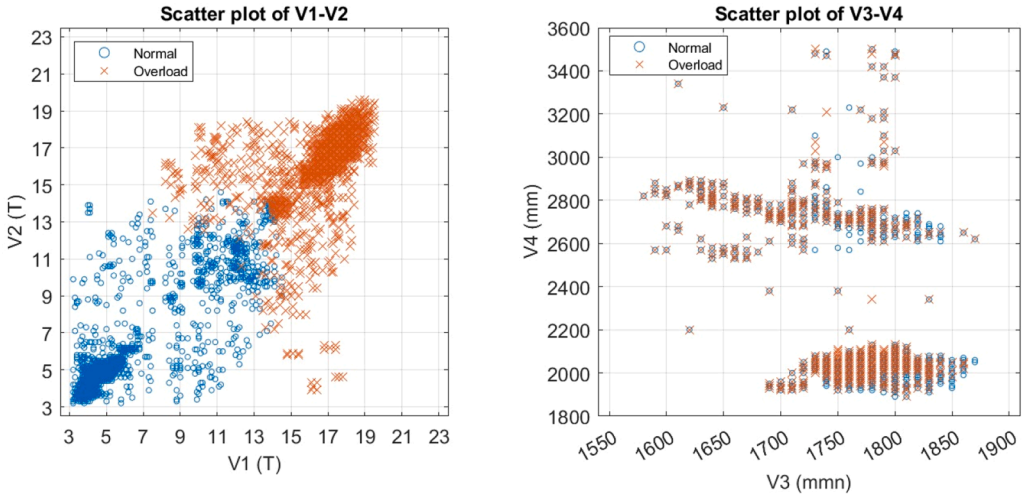


**Fig. 19.** Variable distribution of subset WIM_ALC_6 (OP=50 %).

**Table 9**
Cross validation results of NN models with varying overload proportions.

| WIM_ALC subsets | Original | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|---|
| OP (%) | 5 % | 10 % | 15 % | 20 % | 30 % | 40 % | 50 % |
| Accuracy | 99.92 % | 99.89 % | 99.84 % | 99.91 % | 99.80 % | 99.71 % | 99.78 % |
| F1-Score | 99.23 % | 99.45 % | 99.46 % | 99.76 % | 99.66 % | 99.64 % | 99.78 % |
| Precision | 99.36 % | 99.40 % | 99.61 % | 99.75 % | 99.71 % | 99.60 % | 99.77 % |
| Recall | 99.11 % | 99.50 % | 99.31 % | 99.77 % | 99.60 % | 99.67 % | 99.79 % |
| AUC | 99.81 % | 99.84 % | 99.79 % | 99.94 % | 99.91 % | 99.95 % | 99.94 % |

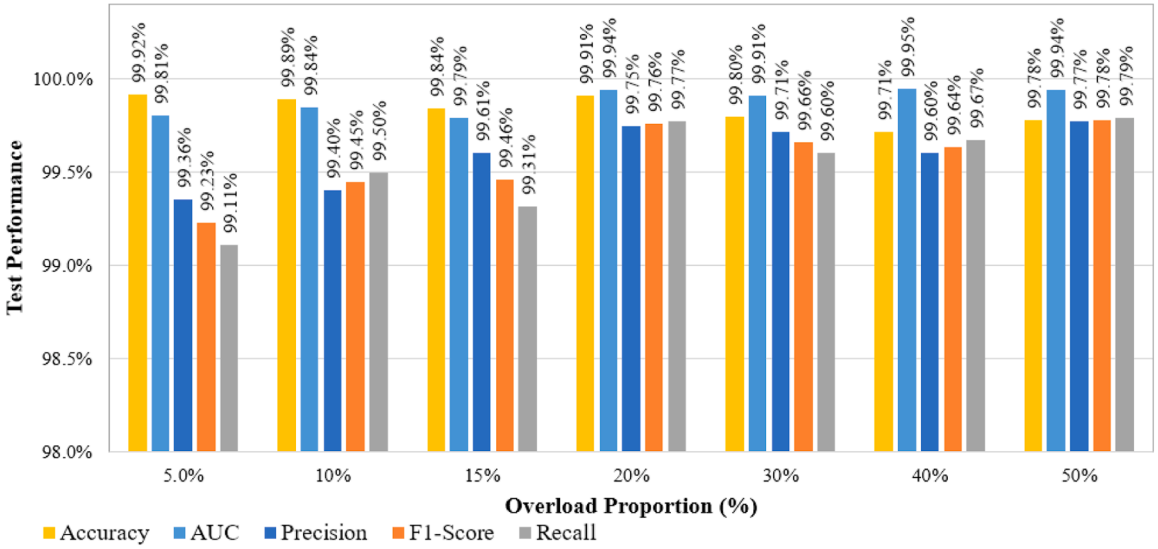


**Fig. 20.** NN model performance results with varying overload proportions.

overall accuracy. These results confirm the feasibility of the developed NN model for assessing overload conditions in railway bridges using WIM_ALC data.

### 3.2.3. *Optimal overload proportion*

This section investigates the influence of overload data proportion to the accuracy of model prediction. The performance metric results obtained in previous section can be satisfactory within current ML-based classification practices [33,34]. However, since misclassification of overloaded ALCs can lead to unnoticed damage in affected railway bridges, it is necessary to maximise the recall metric (minimise false negative detections). It is also desirable to keep the precision rate as high as possible to reduce unnecessary line closure due to false positive detection. Detailed analysis results in Table 7 and Fig. 17 reveal noticeably unstable recall and precision values among the five tests, particularly the significant low recall values of 98.49 % in Test 3 and precision value of 98.87 % in Test 4 compared to other tests. This can likely be attributed to the imbalance between the normal and overload classes in the dataset. Therefore, it is necessary to investigate the effect of overload data proportion on model classification accuracy. Toward this end, different overload proportions (OP) ranging from 10 % to 50 % were assigned to the current WIM_ALC dataset to find the equivalent

**Table 10**
Artificially created axle load combination datasets.

| Dataset | Number of Variable | | | | Data size | Normal ALC cases | Overload ALC cases | Overload proportion |
|---|---|---|---|---|---|---|---|---|
| | V1 | V2 | V3 | V4 | | | | |
| **SYN_ALC_1k** | 7 | 7 | 4 | 7 | **1372** | 1221 | 151 | 11.0 % |
| **SYN_ALC_5k** | 10 | 10 | 5 | 10 | **5000** | 4459 | 541 | 10.8 % |
| **SYN_ALC_10k** | 14 | 14 | 5 | 10 | **9800** | 8602 | 1198 | 12.2 % |
| **SYN_ALC_20k** | 18 | 18 | 5 | 12 | **19,440** | 15658 | 3782 | 19.5 % |
| **SYN_ALC_50k** | 23 | 23 | 7 | 14 | **51,842** | 37271 | 14571 | 28.1 % |
| **SYN_ALC_100k** | 32 | 32 | 7 | 14 | **100,352** | 70,764 | 29,588 | 29.5 % |
| **SYN_ALC_250k** | 43 | 43 | 8 | 17 | **251,464** | 159,282 | 92,182 | 36.7 % |
| **SYN_ALC_500k** | 52 | 52 | 11 | 17 | **539,784** | 349,927 | 155,721 | 30.8 % |
| **SYN_ALC_1m** | 63 | 63 | 16 | 17 | **1079,568** | 682,426 | 397,142 | 36.8 % |
| **SYN_ALC_5m** | 68 | 68 | 21 | 52 | **5049,408** | 3519,977 | 1529,431 | 30.3 % |

allowable bending moment $M_{allow\text{-}live}$ thresholds. This could be done conveniently by interpolating the CDF function created in Section 3.1.3 with results are shown in Table 8. Accordingly, six new WIM_ALC datasets were created following the data labelling process described in Section 3.1.3 applied for each of the six created $M_{allow\text{-}live}$ thresholds.

Distributions of the new ALC variables are depicted in Fig. 18 and Fig. 19, illustrating the increase in the presence of overload data in the datasets. Finally, the datasets were divided into training and testing subsets for training and testing with the NN model using cross-validation techniques, following the same procedure used in Section 3.2.2. The average performance results for each dataset are summarised in Table 9 and plotted in Fig. 20. It is evident that increasing overload proportion contributes to the improvement of the model classification accuracy. In addition, a 20 % overload proportion or higher provides satisfactory balance among Precision, Recall and F1-score indices, with significant improvement in the accuracy levels. These findings inform the creation of training datasets for ML-based RBSO classification model development for practical application, as demonstrated in the next section.

## 4. Final development of RBSO assessment model using synthetic ALC data

Due to the shortage of actual axle overload events, it is essential to artificially create overloaded data for the final model development. Building upon the above preliminary investigation, this section shows how such a dataset, here called synthetic axle load combination (SYN_ALC) datasets, is generated and used to develop a comprehensive RBSO classification model. By expanding the range of ALC variables, more realistic structural overload thresholds ($M_{allow\text{-}live}$) can be applied, resulting in a robust model capable of both binary and multi-class classifications.

### 4.1. Creation of synthetic axle load combinations

Based on the insights gained from analysing the real WIM_ALC dataset in Section 3.1.2, this section presents a method for artificially generating synthetic axle load combinations. Initially, typical distribution ranges of the four ALC variables (Fig. 8) were determined. Subsequently, artificial ALC datasets were generated using the elemental (variable) combination method. To ensure comprehensive model development, this methodology prioritises the inclusion of diverse ranges of axle load and spacing combinations, encompassing both typical operational scenarios and extreme overload conditions that, while unlikely in practice, are essential for model robustness. These extreme cases include scenarios such as consecutive overloaded wagons (e.g., V1 =V2 =24t) or an empty wagon (possibly unloaded at intermediate stations) followed by an overloaded one (e.g., V1 =3t, V2 =24t). Even though such events may be rare, the model must be capable of accurately identifying and responding to them if they were to occur.

Fig. 9 and Fig. 10 (Section 3.1.2) show that most of the recorded WIM axle weights (V1, V2) fall within the normal range from 3 t to 19.5 t. To accommodate a reasonable upper limit for the axle weight variables, reference was made to KiwiRail's overload guidelines [28], which categorize axle overload levels as Medium (>19.8 t and ≤ 20.7 t), High (>20.7 t and ≤ 21.6 t), and Extreme (>21.6 t), as summarised in Table 1. To encompass all potential overload scenarios, V1 and V2 were reasonably expanded to range from 3 t to 24 t, with 24 t representing an approximate 10 % margin above the Extreme axle overload threshold.

Similarly, from analysing Fig. 11, the bogie axle spacing V3 variable

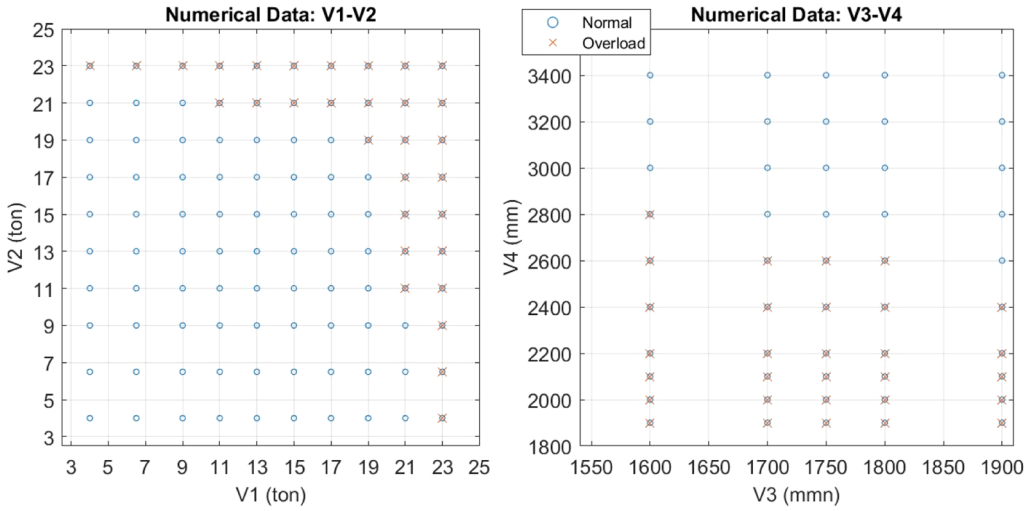


**Fig. 21.** Variable distribution: SYN_ALC_5k.

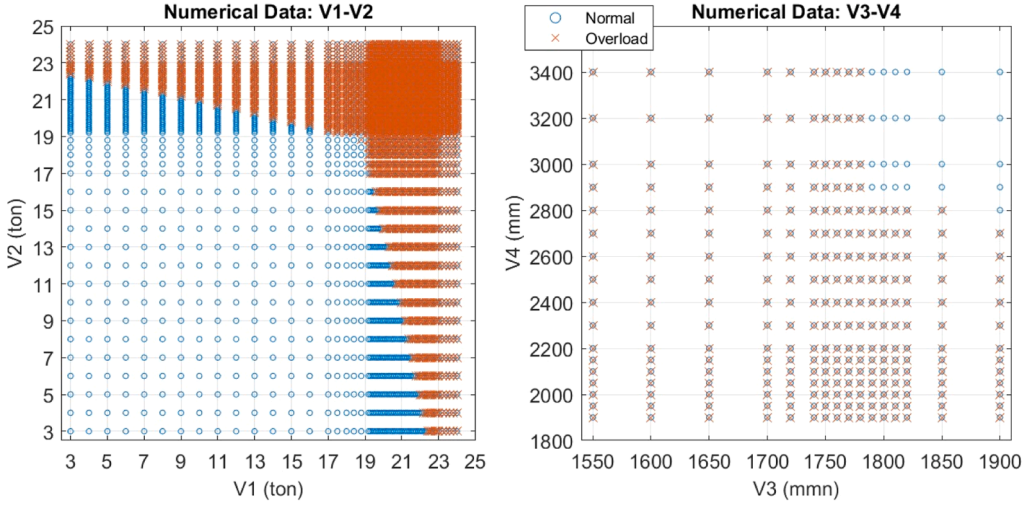


**Fig. 22.** Variable distribution: SYN_ALC_1m.

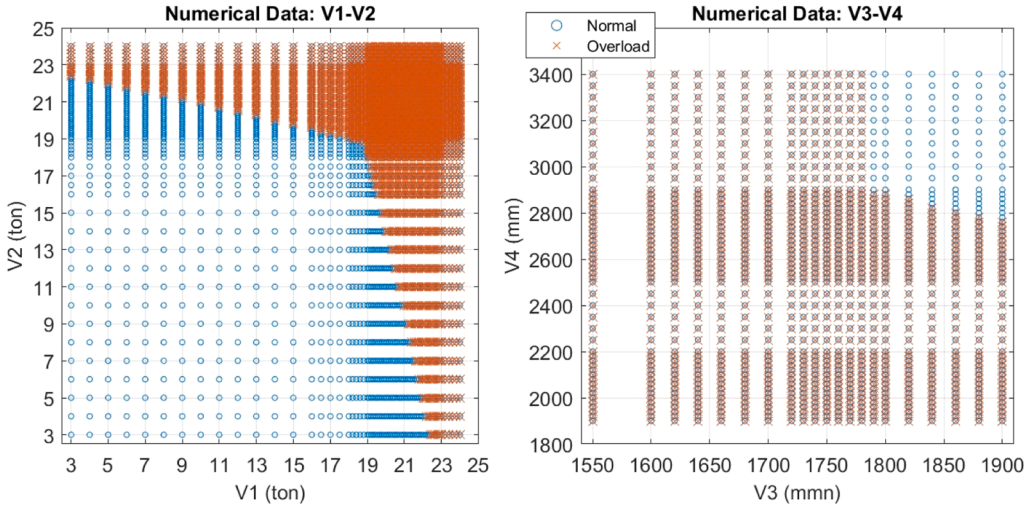


**Fig. 23.** Variable distribution: SYN_ALC_5m.

can be generated in the range of [1600÷1900] mm, with higher concentration of around 1770 mm. The wagon spacing V4 variable was created in the range of [1900÷3400] mm, with higher density in the [1900÷2100] mm and [2500÷2900] mm ranges. V4 values exceeding 3400 mm were excluded due to their negligible influence on the $M_{max}$ results of 6 m long bridges.

The elemental combination method was employed to create ALCs by combining the four generated variables. To vary the dataset size, the density or sparsity of the variables was suitably adjusted, resulting in datasets ranging from over one thousand (SYN_ALC_1k) to five million (SYN_ALC_5m) instances or even more if necessary (Table 10, Fig. 21 to Fig. 23). Preliminary analyses indicate that training the model to accurately differentiate ALCs near the subtle boundary between normal and overload states necessitates a higher density of V1 and V2 data points exceeding 17 t. Conversely, as all ALCs with both V1 and V2 values below 17 t are classified as normal, a sparse distribution of these variables within the 3 t to 17 t range can be employed without compromising model performance (Fig. 22, Fig. 23). The data size was then determined by multiplying the number of variables within their respective ranges, as shown in Table 10. The synthetic datasets generated in this manner mimic the gradual accumulation of WIM measurement data over time. Again, this approach has the advantage of systematically generating variable distributions that represent the full range of potential ALCs, thereby enhancing the model's robustness in interpreting and classifying future wagon ALC data.

### 4.2. SYN_ALC data label

For each ALC in the SYN_ALC datasets, $M_{max}$ values were then calculated for a specified 6-meter-long single-span bridge and used to label the data. To evaluate the influence of data size on classification accuracy and to determine the optimal data size, the datasets were initially labelled with a single structural overload level, which results in binary classes: 'normal' and 'overload'. For this, the first axle overload threshold (V1 = V2 = 19.8 T) was applied to a typical Cooper E series configuration with axle spacing (V3) of 1.73 m and bogie spacing (V4) of 2.00 m [3,35], yielding a $M_{allow\text{-}live}$ of 512.01 kNm. It is important to note that this $M_{allow\text{-}live}$ is an assumed threshold used for demonstrating the methodology in this research. In practical applications, it should be determined through a structural health monitoring program, as previously discussed in Section 2.1.1. Accordingly, ALCs were labelled as '0' ('normal') if $M_{max} \leq M_{allow\text{-}live}$, or '1' ('overload') if $M_{max} > M_{allow\text{-}live}$. As shown in Table 10, most SYN_ALC datasets contain over 20 % overload proportions, ensuring balanced classification results among the

**Table 11**
Examples of the SYN_ALC dataset for training and testing.

| V1 | V2 | V3 | V4 | True_Class | $M_{max}$ |
|---|---|---|---|---|---|
| 9 | 17 | 1650 | 2500 | 0 | 385.09 |
| 17 | 21.3 | 1900 | 3000 | 0 | 444.07 |
| 17 | 21.3 | 1900 | 3200 | 0 | 444.07 |
| 17 | 21.3 | 1900 | 3400 | 0 | 444.07 |
| 17 | 21.5 | 1650 | 1950 | 1 | 546.33 |
| 20.3 | 22.3 | 1600 | 2700 | 1 | 513.45 |
| 20.3 | 22.3 | 1600 | 2800 | 0 | 504.01 |
| 21.5 | 14 | 1650 | 3400 | 0 | 470.7 |
| 21.5 | 14 | 1700 | 1900 | 1 | 529.71 |

performance indices, as suggested by preliminary study in Section 3.2.3. Examples of the SYN_ALC instances are presented in Table 11, which are similar to the WIM_ALC data presented in Table 4.

### 4.3. Selection of SYN_ALC data size for RBSO classification

The SYN_ALC datasets generated in Section 4.2 were fed into the selected Wide Neural Network model for training and testing to find the optimal dataset size for achieving highly accurate RBSO classification. The trained models were then evaluated using a test dataset comprising 10,538 instances from the WIM_ALC dataset (Sections 3.1) and 5,049 instances, or 1 %, from the SYN_ALC_5m dataset. To ensure the test data was distinct from the training SYN_ALC datasets, a random noise of 5 % was added separately to each variable in the 5049 SYN_ALCs of the test dataset. Applying the selected $M_{allow\text{-}live} = 512.01$ kNm threshold to the test data resulted in a total of 15587 ALCs, including 14,044 normal ALCs (from both WIM_ALC and SYN_ALC data) and 1543 overload ALCs (9.9 % of the data size, exclusively from SYN_ALC data), with a diverse range of variables as depicted in Fig. 24.

The training-testing procedure was conducted at least 10 times for each synthetic dataset to avoid outliers and to obtain average performance indices with minimal standard deviations. The subsequent test performance results are summarised in Table 12 and Fig. 25, which show that both Accuracy and F1-Score indices are proportional to the training data size. F1-Score is shown to be more sensitive to changes in training data size, exhibiting a wider range of 91.17 % to 99.87 % compared to the relatively stable Accuracy index, which varies from 98.35 % to 99.97 %. In addition, the F1-Score increases rapidly from 91.17 % to 99.42 % as training data size grows from 1000 to 100,000, but then stabilizes around 99.84 % for datasets of one million or more. Based on these findings, a dataset size of one million is sufficient for training a highly accurate NN classification model. Consequently, the SYN_ALC_1m was selected for further model performance analyses. This choice is supported by its reasonable training time of only 13.5 min compared to the hour required for the 5-million dataset (Table 12).

Examples of detailed prediction results of the NN models trained with the SYN_ALC_1m dataset are presented in Table 13. The results demonstrate the model's robust classification capabilities, with a median of only 3.6 positively and 2.4 negatively misclassified data samples (out of over 15,000 testing samples) per experiment. Compared to the previous results using WIM_ALC dataset in Section 3.2.2 (Table 7), significant improvements were observed in Precision, Recall, and F1-score. With a Precision of 99.81 %, Recall of 99.87 %, and F1-score of 99.84 %, the model demonstrated exceptional accuracy in detecting both classes, particularly its ability to correctly identify positive cases with minimal false negatives. This meets the critical requirement in railway bridge safety management, where minimising false negative structural overload detections is crucial to prevent serious incidents.

Table 14 summarizes all misclassified ALC samples accumulated from different testing trials. The analysis reveals that these misclassifications occurred in ALC samples with $M_{max}$ values marginally differing from the allowable value of 512.01 kNm. This indicates that the developed model can effectively identify all clear-cut ALCs that pose a threat to bridge safety. However, it may struggle to differentiate ALCs that fall within the subtle limit state between normal and overload,

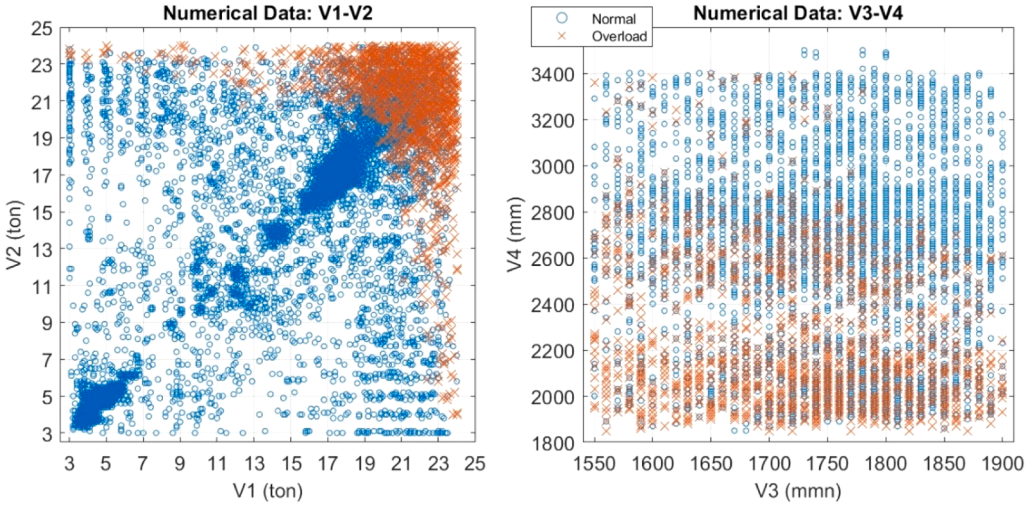


**Fig. 24.** Characterization of the test dataset with noise added.

**Table 12**
Test results of NN models using different SYN_ALC datasets.

| Train Datasets | 1k | 5k | 10k | 20k | 50k | 100k | 250k | 500k | 1 m | 5 m |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | 98.35 % | 98.91 % | 99.29 % | 99.65 % | 99.83 % | 99.88 % | 99.91 % | 99.95 % | 99.97 % | 99.97 % |
| F1-Score | 91.17 % | 94.35 % | 96.29 % | 98.23 % | 99.15 % | 99.42 % | 99.55 % | 99.74 % | 99.84 % | 99.87 % |
| Std.Dev(F1) | 0.74 % | 0.47 % | 0.48 % | 0.17 % | 0.13 % | 0.09 % | 0.06 % | 0.06 % | 0.07 % | 0.06 % |
| Training time (min.)* | < 0.10 | < 0.10 | < 0.10 | < 0.10 | 0.20 | 1.00 | 2.90 | 6.85 | 13.5 | 59.4 |

Note:
* Training performed by a NVIDIA DGX Station V100 system

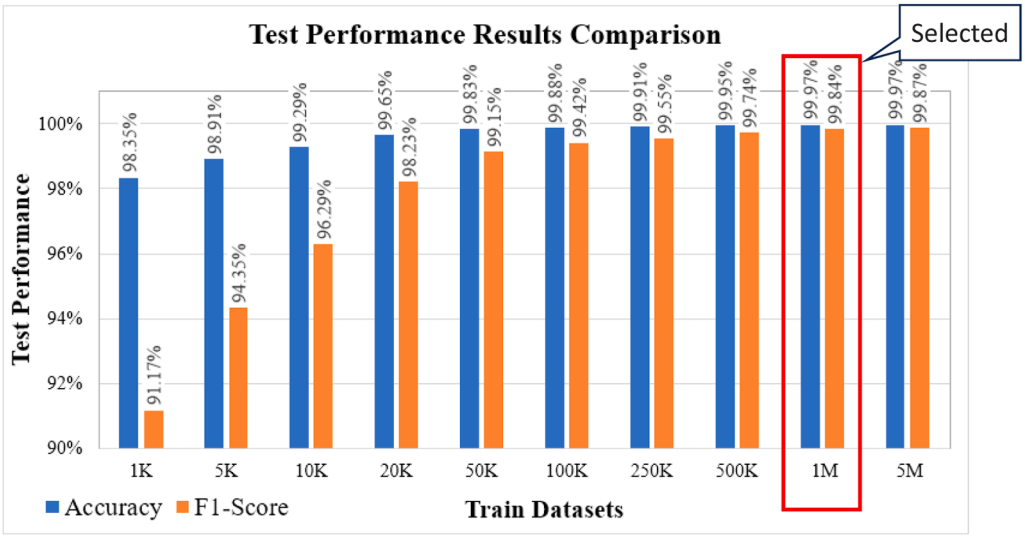


**Fig. 25.** Performance of NN models with different SYN_ALC training datasets.

**Table 13**
Prediction results of the NN models trained with SYN_ALC_1m dataset.

| Model No. | TP | FP | TN | FN | Accuracy (%) | Precision (%) | Recall (%) | F1-Score (%) |
|---|---|---|---|---|---|---|---|---|
| 1 | 1541 | 3 | 14041 | 2 | 99.97 | 99.81 | 99.87 | 99.84 |
| 2 | 1541 | 3 | 14041 | 2 | 99.97 | 99.81 | 99.87 | 99.84 |
| 3 | 1540 | 5 | 14039 | 3 | 99.95 | 99.68 | 99.81 | 99.74 |
| 4 | 1542 | 3 | 14041 | 1 | 99.97 | 99.81 | 99.94 | 99.87 |
| 5 | 1539 | 4 | 14040 | 4 | 99.95 | 99.74 | 99.74 | 99.74 |
| **Medians** | | **3.6** | | **2.4** | **99.97** | **99.81** | **99.87** | **99.84** |

**Table 14**
Collection of false detection cases.

| V1 (t) | V2 (t) | V3 (mm) | V4 (mm) | True_Class | Prediction class | False type | Mmax (kNm) | ΔM (kNm) |
|---|---|---|---|---|---|---|---|---|
| 17.7 | 23.3 | 1690 | 2780 | 1 | 0 | FN | 512.29 | 0.28 |
| 19.2 | 22.9 | 1750 | 2640 | 0 | 1 | FP | 511.99 | −0.02 |
| 20.3 | 20.2 | 1700 | 2160 | 0 | 1 | FP | 511.94 | −0.07 |
| 19.4 | 20.2 | 1850 | 1940 | 1 | 0 | FN | 512.05 | 0.04 |
| 22.5 | 18.6 | 1880 | 2370 | 1 | 0 | FN | 512.19 | 0.18 |
| 22.4 | 11.7 | 1590 | 2530 | 0 | 1 | FP | 511.77 | −0.24 |
| 22.4 | 18.9 | 1750 | 2520 | 0 | 1 | FP | 511.91 | −0.10 |
| 23.2 | 19.7 | 1610 | 2940 | 0 | 1 | FP | 511.85 | −0.16 |
| 23.0 | 22.3 | 1710 | 2770 | 1 | 0 | FN | 512.15 | 0.14 |
| 18.8 | 22.4 | 1630 | 2670 | 0 | 1 | FP | 511.77 | −0.24 |

**Table 15**
Multiple RBSO Level Definitions.

| RBSO classes | Explanations | $M_{allow\text{-}live}$ (kNm) | Data label | Class proportion |
|---|---|---|---|---|
| Normal | Normal threshold | ≤ 512.01 | '0' | 63.2 % |
| Medium | Overload of 0 % to 10 % | 512.01 ÷ 535.29 | '1' | 13.8 % |
| High | Overload of 10 % to 15 % | 535.29 ÷ 558.56 | '2' | 10.6 % |
| Extreme | Overload of greater than 15 % | > 558.56 | '3' | 12.4 % |

producing some false positive and negative classifications. However, in practice, the false negative rate can be reduced by conservatively setting a slightly higher $M_{allow\text{-}live}$ threshold during training.

### *4.4. Performance of the NN Model in Multi-level Structural Overload Classification*

Previous investigations primarily focused on a single overload level applied uniformly to all bridges of the same span. However, in real-world scenarios, the load-bearing capacity can vary among the bridges due to many factors, such as structural upgrades or damage. This necessitates the ability of the classification models in classifying multiple overload levels. To evaluate the NN model's capacity in this regard, the previously selected SYN_ALC_1m dataset was assumed to be categorised into one normal and three RBSO classes, with their proportions and labels summarized in Table 15. This classification establishes the first overload level at a $M_{allow\text{-}live}$ of 512.01 kNm (as used in Section 4.3), followed by two additional overload levels of 10 % ($M_{allow\text{-}live}$ = 535.29 kNm) and 15 % ($M_{allow\text{-}live}$ = 558.56 kNm) higher.

Once the dataset was relabelled, it was fed into the NN model for training and testing using the hold-out cross-validation technique, where 80 % of the ALCs was used for training and the remaining 20 % was held back for testing. This process was then repeated ten times, each

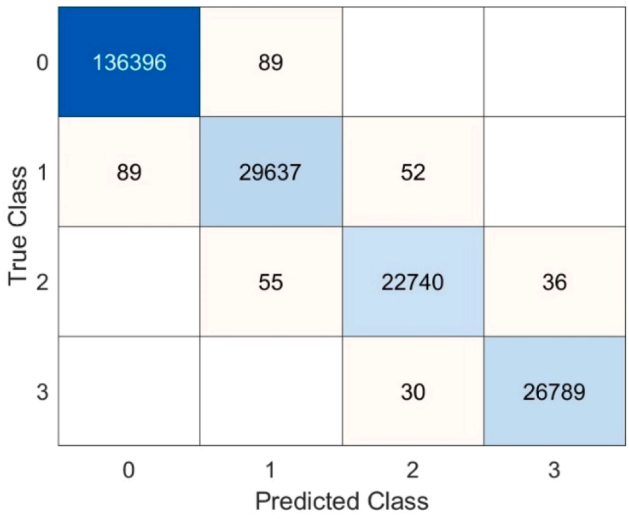


**Fig. 26.** A typical confusion matrix of multi-level RBSO classification.

with a different random training-testing partitioning. Fig. 26 shows a typical confusion matrix of the testing results for multi-class classification. Based on the testing output, the four test performance indices were calculated for each class, and the average results are summarized in Fig. 27. The results indicate a clear distinction among the four classes, with false detections occurring only between consecutive overload levels. For instance, the Normal class ("0") was only misclassified as Medium ("1"), while no misclassifications of Normal instances occurred in High ("2") or Extreme ("3"). Importantly, the developed classification model demonstrated exceptional capability in identifying and differentiating various structural overload levels, with well-balanced and highly accurate results (approximately 99.5 % or higher) across all four performance indicators. The model's success in this capacity enables bridges with high load-carrying capacities to be exempted from unnecessary structural damage inspections during lower-level structural overload events, making it ideally suited for real-world RBSO applications.

## 5. Conclusion

A supervised machine learning (ML) approach for assessing structural overload in railway bridges using real and synthetic weigh-in-motion (WIM) data was proposed and presented in this paper. A two-stage model development process was employed. Using a dataset derived from real WIM measurements, the research initially focused on extracting and analysing key characteristics of train axle load combination (ALC) variables, selecting appropriate ML models, and investigating the impact of overload data proportion on model performance. Subsequently, a comprehensive railway bridge structural overload (RBSO) classification model was developed utilising synthetic ALC datasets generated based on real WIM dataset analysis.

The initial model development stage explored the feasibility of applying supervised ML approach for RBSO assessment using a real WIM_ALC dataset containing over 10,000 instances. A thorough evaluation revealed that the Neural Network (NN) outperformed other machine learning (ML) models, achieving a nearly perfect validation accuracy of 99.90 %. The selected NN model demonstrated impressive classification capabilities, with a precision rate of 99.36 %, recall of 99.11 %, and a harmonized F1-score of 99.23 % in testing unseen data. A parametric study indicated that maintaining an overload percentage of 20 % or higher in the training dataset is crucial for achieving both high accuracy and optimal balance among the performance indices.

However, obtaining such a proportion of overload events solely from real WIM data is challenging due to their scarcity. To address this data imbalanced issue, the final model development stage introduced a method for artificially generating synthetic ALC datasets, encompassing both typical operational scenarios and extreme overload conditions. This augmentation process resulted in various SYN_ALC datasets with varying data sizes, ranging from over one thousand to five million instances or more. A comprehensive study on the influence of data size on classification accuracy demonstrated that a dataset size of one million is sufficient for training a highly accurate NN classification model. This model achieved a substantial increase in performance metrics, with the precision rate rising to 99.81 %, recall to 99.87 %, and F1-score to 99.84 % for single overload level classification.

Finally, the NN model's capability was examined under diverse RBSO rating criteria. The results demonstrated the model's exceptional ability to identify and differentiate various structural overload levels, consistently achieving well-balanced and highly accurate results (99.5 % or higher) across all four performance indicators. This proves the model's applicability in providing instant structural overload assessments for bridges with varying load-bearing capacities.

In summary, this paper contributed a robust ML-based tool for automated structural overload assessment in railway bridges, offering

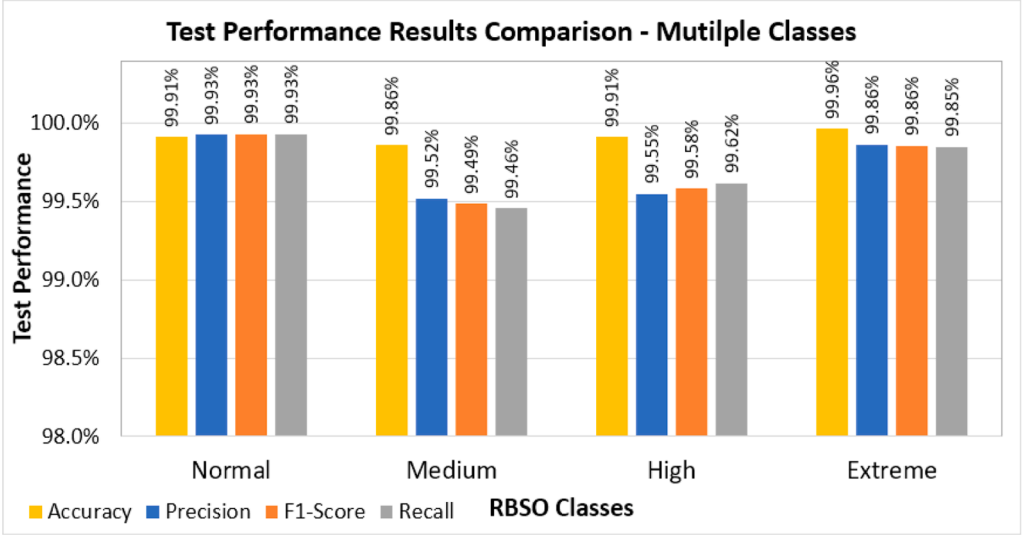


**Fig. 27.** Performance of the NN model in multi-level RBSO classification.

efficiency and accuracy compared to conventional methods. The findings lay the groundwork for future research utilising more extensive WIM datasets, with larger number of variables and varying span length, or incorporating other synthetic data generation techniques, such as the generative adversarial networks, or addressing more complex rail bridge rating criteria. Additionally, the applicability of the proposed methodology can be extended beyond the current scope to other bridge types, including multiple-span and truss railway bridges. Lastly, while this study has focused on bending moment response, future research can incorporate other structural response parameters to enrich the overload assessment applications in bridges.

## CRediT authorship contribution statement

**N.T.Le:** Writing – original draft, Writing – review & editing, Methodology, Investigation, Formal analysis, Visualization, Validation, Software, Conceptualization. **M.Keenan:** Writing – original draft, Writing – review & editing, Methodology, Investigation, Formal analysis, Visualization, Software, Conceptualization. **A.Nguyen:** Writing – review & editing, Supervision, Software, Resources, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Conceptualization. **S.Ghazvineh:** Writing – review & editing, Investigation, Visualization. **Y.Yu:** Writing – review & editing, Investigation, Visualization. **J.Li:** Writing – review & editing, Investigation, Visualization. **A.Manalo:** Writing – review & editing, Investigation, Visualization.

## Declaration of Competing Interest

The authors declare that they have no known competing financial and non-financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

The authors wish to thank KiwiRail, the state-owned corporation that handles rail operations in New Zealand, for permission to extract and use the authentic WIM measurement data in this research.

## References

[1] Mithraratne N, Lifetime liabilities of land transport using road and rail infrastructure. 2011.
[2] Žnidarič A, Pakrashi V, O'Brien E, O'Connor A. A review of road structure data in six European countries. Proc Inst Civ Eng-Urban Des Plan 2011;164(4):225–32.
[3] Nassif H , Ozbay K , Iyer S , Su D , Lou P , Kara E , Capers Jr H , Valeo M , Elimination of weight restriction on Amtrak, NJ Transit, and Conrail lines. 2012.
[4] Lydon M, Taylor SE, Robinson D, Mufti A, Brien E. Recent developments in bridge weigh in motion (B-WIM). J Civ Struct Health Monit 2016;6:69–81.
[5] Ye XW, Ni YQ, Wong KY, Ko JM. Statistical analysis of stress spectra for fatigue life assessment of steel bridges with structural health monitoring data. Eng Struct 2012;45:166–76.
[6] Maes K, Van Meerbeeck L, Reynders EPB, Lombaert G. Validation of vibration-based structural health monitoring on retrofitted railway bridge KW51. Mech Syst Signal Process 2022;165:108380.
[7] Le, N. T., Nguyen, A., Chan, T. H. T., Thambiratnam, D. P., Damage Identification in Large-Scale Bridge Girders Using Output-Only Modal Flexibility–Based Deflections and Span-Similar Virtual Beam Models, *Structural Control and Health Monitoring*, 2024, 4087831, 32 pages, 2024. https://doi.org/10.1155/2024/4087831.
[8] Zhang G, Liu Y, Liu J, Lan S, Yang J. Causes and statistical characteristics of bridge failures: a review. J Traffic Transp Eng 2022;9(3):388–406.
[9] Yu Y, Cai C, Deng L. State-of-the-art review on bridge weigh-in-motion technology. Adv Struct Eng 2016;19(9):1514–30.
[10] Lee CE and Garner JE , Collection and analysis of augmented weigh-in-motion data. 1996, University of Texas at Austin. Center for Transportation Research.
[11] Van der Spuy P, Lenner R, de Wet T, Caprani C. Multiple lane reduction factors based on multiple lane weigh in motion data. Structures 2019;20:543–9.
[12] Sujon M, Dai F. Application of weigh-in-motion technologies for pavement and bridge response monitoring: state-of-the-art review. Autom Constr 2021;130:103844.
[13] Bushman R and Pratt AJ , Weigh in motion technology–economics and performance. in Presentation on the North American Travel Monitoring Exhibition and Conference (NATMEC). Charlotte, North Carolina. 1998.
[14] Moses F, Weigh-in-Motion System Using Instrumented Bridges. 1979. **105**(3): p. 233–249.
[15] Carraro F, Gonçalves MS, Lopez RH, Miguel LFF, Valente AM. Weight estimation on static B-WIM algorithms: A comparative study. Eng Struct 2019;198:109463.
[16] Hajializadeh D , Žnidarič A , Kalin J , OBrien EJ , Development and Testing of a Railway Bridge Weigh-in-Motion System. 2020. **10**(14): p. 4708.
[17] Deng Y, Zhang M, Feng D-M, Li A-Q. Predicting fatigue damage of highway suspension bridge hangers using weigh-in-motion data and machine learning. Struct Infrastruct Eng 2021;17(2):233–48.
[18] Xiao X, Pi D, Zhu Q. A bridge weigh-in-motion algorithm for fast-passing railway freight vehicles considering bridge-vehicle interaction. Mech Syst Signal Process 2022;181:109493.
[19] Szinyéri B, Kővári B, Völgyi I, Kollár D, Joó AL. A strain gauge-based Bridge Weigh-In-Motion system using deep learning. Eng Struct 2023;277:115472.
[20] Pimentel R, Ribeiro D, Matos L, Mosleh A, Calçada R. Bridge Weigh-in-Motion system for the identification of train loads using fiber-optic technology. Structures 2021;30:1056–70.
[21] Wei Y-T, Yi T-H, Yang D-H, Li H-N. Bridge Damage Localization Using Axle Weight Time History Data Obtained through a Bridge Weigh-in-Motion System. J Perform Constr Facil 2021;35(5):04021065.
[22] Mosleh A , Costa PA , Calçada R , A new strategy to estimate static loads for the dynamic weighing in motion of railway vehicles. 2020. **234**(2): p. 183–200.
[23] Pau A, Vestroni F. Weigh-in-motion of train loads based on measurements of rail strains. Struct Control Health Monit 2021;28(11):e2818.
[24] Pintão B , Mosleh A , Vale C , Montenegro P , Costa P , Development and Validation of a Weigh-in-Motion Methodology for Railway Tracks. 2022. **22**(5): p. 1976.
[25] Yannis G, Antoniou C. Integration of weigh-in-motion technologies in road infrastructure management. ITE J 2005;75(1):39–43.
[26] Engineering New Zealand. Coupled in motion train weighing. 2022; Available from: ⟨https://www.engineeringnz.org/programmes/heritage/heritage-records/coupled-in-motion-train-weighing/⟩.
[27] KiwiRail, Structures Task Instruction: Structural Evaluation of Overload. 2021.
[28] KiwiRail, Structures Standard: Monitoring and Inspection of Overloaded Structures. 2016.
[29] AREMA, Manual of Railway Engineering. 2009, American Railway Engineering and Maintenance-of-Way Association.
[30] KiwiRail, Bridge Moving Load Calculation. Structural Requirements and Background to Process, K.i. document, Editor. 2011.
[31] Thai H-T. Machine learning for structural engineering: a state-of-the-art review. Structures 2022;38:448–91.
[32] Trappenberg TP. Fundamentals of machine learning. Oxford University Press; 2019.
[33] Nguyen A, Gharehbaghi V, Le NT, Sterling L, Chaudhry UI, Crawford S. ASR crack identification in bridges using deep learning and texture analysis. Structures. Elsevier,; 2023.
[34] Nguyen A , Chianese RR , Gharehbaghi VR , Perera R , Low T, Aravinthan T , Yu Y , Samali B, Guan H , Khuc T , Robustness of Deep Transfer Learning-Based Crack Detection against Uncertainty in Hyperparameter Tuning and Input Data. 2022.
[35] Leighty CA, Laman JA, Gittings ‡ GL. Heavy axle study: impact of higher rail car weight limits on short-line railroad bridge structures. Civ Eng Environ Syst 2004;21(2):91–104.