2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10340243

Identification of injured elements in computational models of spinal cord injury using machine learning *

Cesar Jimenez, Carolyn J. Sparrey, and Mohammad Narimani, *Member, IEEE*

***Abstract*— The purpose of this study was to use machine learning (ML) algorithms to identify tissue damage based on the mechanical outputs of computational models of spinal cord injury (SCI). Three datasets corresponding to gray matter, white matter, and the combination of gray and white matter tissues were used to train the models. These datasets were built from the comparison of histological images taken from SCI experiments in non-human primates and corresponding subject-specific finite element (FE) models. Four ML algorithms were evaluated and compared using cross-validation and the area under the receiver operating characteristic curve (AUC). After hyperparameter tuning, the AUC mean values for the algorithms ranged between 0.79 and 0.82, with a standard deviation no greater than 0.02. The findings of this study also showed that k-nearest neighbors and logistic regression algorithms were better at identifying injured elements than support vector machines and decision trees. Additionally, depending on the evaluated dataset, the mean values of other performance metrics, such as precision and recall, varied between algorithms. These initial results suggest that different algorithms might be more sensitive to the skewed distribution of classes in the studied datasets, and that identifying damage independently or simultaneously in the gray and white matter tissues might require a better definition of relevant features and the use of different ML algorithms. These approaches will contribute to improving the current understanding of the relationship between mechanical loading and tissue damage during SCI and will have implications for the development of prevention strategies for this condition.**

***Clinical Relevance*— Linking FE model predictions of mechanical loading to tissue damage is an essential step for FE models to provide clinically relevant information. Combined with imaging technologies, these models can provide useful insights to predict the extent of damage in animal subjects and guide the decision-making process during treatment planning.**

I. Introduction

Spinal cord injury is triggered by mechanical loading, which causes a series of biological responses resulting in irreversible functional damage to the neurological system [1]. Understanding the relationship between mechanical loading and tissue damage in the spinal cord could be a critical step to anticipating the injury spreading, particularly in animal models, where it is possible to control the mechanical loading conditions [2], [3]. This preliminary insight into the injury outcomes could provide useful information to define mechanical threshold values, that will result in tissue damage. Injury thresholds would establish relevant criteria for the design of protective equipment, and for clinicians to select the most appropriate treatment to implement [2]–[4]. For these reasons, several studies have looked to establish the relationship between mechanical loading and tissue damage in the spinal cord [2]–[5], but it is not yet well defined.

Computational finite element (FE) models of SCI are a complementary and non-invasive approach to further explore the relationship between mechanical loading and tissue damage [2], [3], [5], [6]. Previous studies have employed a combination of computational models and histopathological findings from SCI experiments to gain a closer insight into the distribution of loads in the spinal cord tissue. For example, mechanical outcomes from FE models of rat [4], [5] and non-human primate (NHP) [3] experiments were correlated with observed biological damage in the animals' tissues using statistical methods such as linear and logistic regression. These studies found that there was a stronger correlation between the mechanical features and the damage in the gray matter (GM) than in the white matter (WM) tissue of the spinal cord. However, during an injury, both GM and WM tissues are subjected to mechanical loading, and therefore it is important to accurately identify the potential injury in both tissues.

Current applications of artificial intelligence in SCI research suggest it can find non-obvious correlations between variables [7]–[9]. For instance, ML algorithms have been used in combination with imaging technologies, such as magnetic resonance imaging (MRI), to identify lesions and damage to the spinal cord [7]. Other applications of ML algorithms have used clinical data to predict changes in functional outcomes after treatment [7], [9], and to assess the pain in patients with SCI [8]. These studies leverage the advantages of ML algorithms to understand complex relationships between variables [7]–[9], motivating their exploration in this study. It is hypothesized that using ML algorithms could improve the identification of injured elements in both the GM and WM tissue based on mechanical loading predictions from FE models of SCI. Moreover, training different ML algorithms with this data could provide clarify the correlation between mechanical load and injury outcomes.

II. Methods

In this study, we aimed to improve a previously proposed method for the identification of injury from mechanical tissue loading data. Datasets collected in [3] were used to train four classification ML algorithms from the scikit-learn library [10]: logistic regression (LR), decision trees (DT), support vector machines (SVM), and k-nearest neighbors (KNN). Then, the

* Research supported by NSERC, WestGrid (www.westgrid.ca) and Compute Canada (www.computecanada.ca), and the SFU Community Trust Endowment Fund.

Cesar Jimenez, Carolyn J. Sparrey, and Mohammad Narimani are with Simon Fraser University - Surrey, Galleria 4, Simon Fraser University, 250-13450 102 Avenue, Surrey, BC V3T 0A3 (phone: +1 604-445-4394; e-mail: cjimenez@sfu.ca).

performance of each model was evaluated to select the best one for the identification of injured elements.

A. *Dataset details*

Pre-injury MRI scans taken from three NHP subjects were used to develop subject-specific FE models matched to in vivo experiments (see Fig. 1-B) [3]. The results from the FE models of the spinal cord tissue were segmented into WM and GM elements. The dataset consisted of five mechanical features with the most relevance and correlation with tissue damage [3]: min/max principal logarithmic strain (LEP), logarithmic strain in axonal direction (LEAXON), Tresca stress (TRESCA), and strain energy density (ESEDEN). Structural tissue damage in the spinal cord was observed from cross-sectional histological slices for each subject [11] (Fig. 1-A). Overlaying the histology data on element slices from the computational models, each element was assigned into one of two target classes: injured or healthy.

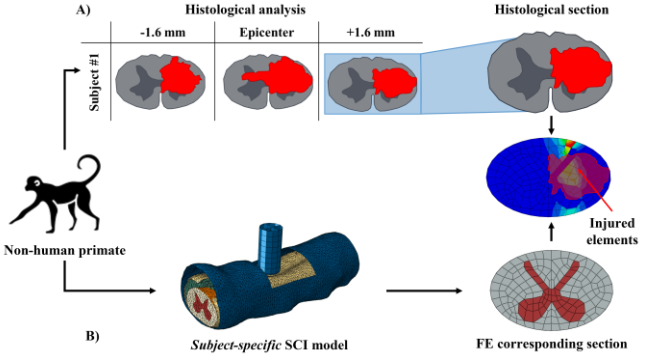

Figure 1. Histology sections and the FE models comparison. Histological analysis performed on the spinal cord of NHP subjects, adapted from [11] (A); an example of subject-specific FE models (B).

After having been assigned to a target class, the elements from the FE models were used to create three datasets across all subjects: GM elements (GM-only), WM elements (WM-only), and combined GM & WM elements (GM&WM) to explore the differences in predicting tissue damage in the spinal cord based on evaluating GM-only or WM-only, or combined tissue elements, GM&WM. The experimental SCI were mild, resulting in more healthy elements in the datasets than injured elements. In addition, there were more WM than GM elements in the dataset due to the tissue distribution in the cervical spinal cord. These uneven distributions of data per tissue type and target value (healthy/injured) were accounted for in the training and implementation of the ML algorithms.

B. *Data pre-processing*

1) *GM-only, WM-only, and GM&WM datasets*

The datasets were imported from comma-separated values files and converted into data frames using Python's pandas library. The datasets were checked for duplicated values and for redundant features, and then split into train, validation, and test datasets with 70%, 20%, and 10% of the original dataset respectively, using the train_test_split_ function [10]. To keep the original class distribution in the split datasets, the 'stratify' parameter was set (Table I). Datasets from GM-only and WM-only were concatenated into a new dataset as GM & WM dataset. To distinguish GM and WM, a new feature column 'TissueType' was included in the dataset.

TABLE I. NUMBER OF HEALTHY (H) AND INJURED (I) ELEMENTS AFTER EACH DATASET SPLIT

Dataset	Train		Validation		Test	
	H	I	H	I	H	I
GM-only	965	553	276	158	138	79
WM-only	3952	1202	1129	344	565	172
GM&WM	4916	1755	1406	502	703	251

C. *Parameter tuning and cross-validation*

To compare the performance of the four ML models on each dataset, a 10-fold cross-validation (CV) was employed using the StratifiedShuffleSplit [10]. This scikit function combined k-fold with shuffle splits to generate randomized sets that preserved the class distribution of the original set. Each algorithm was first fitted with the training portion of the data and then the validation set was used to find the best hyper parameters using the RandomizedSearchCV function [10]. To avoid overfitting during the parameter optimization, an additional 5-fold stratified shuffle split was included in the CV parameter of the RandomizedSearch function. The range of evaluated parameters for each algorithm during the randomized search are described in Table II.

TABLE II. TABLE RANDOMIZED SEARCH PARAMETERS EVALUATED FOR EACH ML ALGORITHM DURING HYPERPARAMETER TUNING

General & algorithms' parameters			
n_iter = 50		scoring = roc_auc	
LR	**DT**	***SVM***	**KNN**
C: loguniform (1e-5, 100)	Criterion: gini, entropy	C: loguniform (1e0, 4e2)	n_neighbour: range(1, 100)
Solver: liblinear, lbfgs, newton-cg	min_samples _split: range (2, 80)	Gamma: auto, scale	Weights: uniform, distance, none
Class_weight: balanced, none	Splitter: best, random	Kernel: rbf, poly, sigmoid	-
-	Class_weight: balanced, none	Class_weight: balanced, none	-

After the hyperparameter tuning, each dataset was split again into new training and test sets, where the validation and test set were combined into a larger evaluation set (30% of the original data) for the CV. The mean and standard deviation (SD) values for the balanced accuracy, precision, recall, F1 score, and AUC metrics were calculated for each algorithm after the 10-fold CV.

III. RESULTS

A. *Parameter tuning and cross-validation*

After 10-fold CV, the AUC mean values and standard deviation for all the algorithms and datasets ranged between 0.79-0.82 and 0.01-0.02, respectively (see Table III).

TABLE III. CV MEAN [SD] AUC VALUES FOR EACH ML ALGORITHM AFTER HYPERPARAMETER TUNING

	LR	**DT**	**SVM**	**KNN**
GM-only	0.805[0.01]	0.808[0.01]	0.808[0.01]	0.811[0.01]
WM-only	0.809[0.01]	0.789[0.02]	0.799[0.02]	0.793[0.02]
GM&WM	0.815[0.01]	0.800[0.02]	0.804[0.02]	0.802[0.02]

The AUC values acquired for the different algorithms were relatively similar for the GM-only dataset (see Table III). Therefore, the F1 score values were used to select the best

performing algorithm, since this metric provides information regarding the trade-off between precision and recall [10]. Based on the new criteria, the KNN algorithm was selected for the GM-only set. LR showed the best classification performance for both the WM-only and GM&WM sets. The combination of parameters that provided these results are listed in Table IV.

TABLE IV. BEST PERFORMING ML ALGORITHMS FOR EACH DATASET WITH THE CORRESPONDING SET OF HYPERPARAMETERS

GM-only	***WM-only***	**GM&WM**
KNN	***LR***	***LR***
n_neighbors: 31	C: 3.756	C: 50.745
weights: uniform	solver: newton-cg	solver: liblinear
-	class_weight: balanced	class_weight: balanced

For a more in-depth comparison of the best algorithms' performance on each dataset, the values of other evaluated metrics: balanced accuracy, precision, recall, and F1 score, were plotted and are shown in Fig. 2.

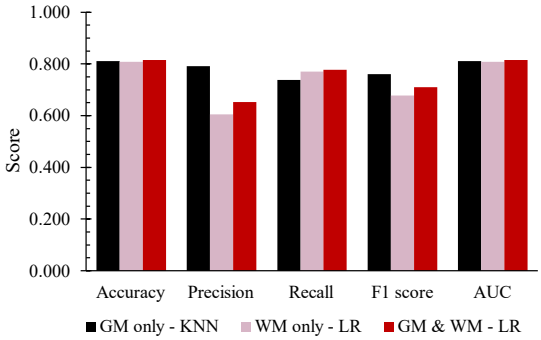

Figure 2. Metrics score of the best performing algorithms for each dataset after CV and hyperparameter tuning, using the 30% test set.

IV. DISCUSSION

A. Parameter tuning and cross-validation

The best AUC scores were achieved by the KNN and LR algorithms. The mean scores and the standard deviation values show a consistent performance during the 10-fold CV for every algorithm. This shows that the hyperparameter tuning process did not overfit the algorithms to the validation data, and the selected parameters contributed to have a robust performance on each iteration of the stratified shuffle split CV.

The best parameters described in Table IV for the ML algorithms help to assess their generalization to new data. With KNN, the 'uniform' weight criteria provided better results than a distance-based approach, implying that the algorithm was less susceptible to overfitting or being influenced by outliers. The LR models for the WM-only and GM&WM sets both worked better with balanced class weights. These results were expected since it is recommended to set this parameter whenever the evaluated dataset is imbalanced [10]. There was also a difference between the 'C' values in both LR algorithms. This parameter, known as the inverse regularization strength [10], controls how much the algorithm will fit the training data. In the WM-only set, the value for 'C' is smaller compared with the one for the GM&WM dataset. Although this might suggest the LR algorithm corresponding to the GM&WM set might be overfitted, it still showed consistent results during the CV.

Despite falling within the narrow range of 0.79 and 0.82, the AUC scores differed depending on the evaluated dataset. These results agree with other biomedical binary classification studies [12], [13], where the performance of ML algorithms was linked to the ability to find patterns in the data [13]. In this study, differences in sample sizes and distribution of classes could further affect the algorithms' sensitivity. For instance, the WM-only set has more than three times the number of samples than the GM-only one (see Table I). At the same time, this smaller sample size increases the percentage of injured elements in this set. Around 36% of the GM-only dataset are injured elements, in contrast with the 23% and the 26% included in the WM-only and GM&WM datasets, respectively. This variance in number of injured elements is related to the structural differences in the GM and WM tissues. The first is a homogeneous and blood vessel-dense tissue [5], while the latter shows anisotropic behavior due to the highly aligned set of axonal fibers [3]; these differences in GM and WM tissue properties affect their mechanical response during injury, and therefore the distribution of damage. These observations highlight the classification challenges of the data and justify the interest in exploring the use of ML to better identify injured sections in the spinal cord tissues.

The unbalanced distribution of injured and healthy elements was one of the reasons the AUC score was selected as the comparison metric. Although in other ML applications the accuracy score determines the classification potential of an algorithm [12], in this study, the skewed class distribution limits this metric's relevance. Different metrics reported in the literature provide a better measure of the classification performance of an algorithm dealing with unbalanced datasets [14]. However, the AUC score was selected as the metric to evaluate, since the results found in [12], [15] show it is robust for both balanced and unbalanced datasets, even those with a greater imbalance compared with these data. Despite the reliability of the AUC metric, the differences in algorithm performance were more evident with other metrics. As shown in Fig. 2, the balanced accuracy and AUC scores were similar between datasets. However, precision scores were significantly lower in the WM-only and GM&WM datasets compared to the GM-only set, indicating there is a greater number of false-positive cases in the classification. These results might be related to the smaller number of injured samples available for those sets and the differences in correlation between mechanical features and the damage in GM or WM tissue found in [3], [5].

Another reason for calculating the AUC values was to compare the results with those reported in [3], where LR was similarly used to investigate the correlation between FE outcomes and tissue damage on the same data. The mean AUC values acquired in [3] ranged between 0.85-0.95 for the GM and 0.72-0.9 for the WM matter. In this study, similar scores were achieved: 0.81 for the GM-only dataset and a range between 0.79-0.81 for the WM-only. Although the acquired range of AUC values are lower than the ones found in [3], one key difference of this study is that collinear features and outlier experimental data were removed from the input samples. In [3], the data of a FE model corresponding to a NHP subject

whose experiment deviated from the expected outcomes was included. The results from this experiment and FE model were significantly different than the ones found in other subjects, therefore those samples were excluded from the input data used in this study. During the data pre-processing, the von-Mises and TRESCA features had a Pearson correlation coefficient of 1, indicating collinearity between the variables, therefore a LR model was fitted to the training data and the feature coefficients were compared; as a result, only the TRESCA feature was included. Ensuring the quality of the data and avoiding additional sources of variability are a fundamental step to obtaining more accurate results, which can allow us to draw more reliable conclusions regarding the relationship between mechanical loading and tissue damage during SCI.

V. Limitations

The available data from the NHP experiments included a limited number of samples from only three subjects for the training and validation of the ML algorithms. A larger sample size could reduce the risk of overfitting and provide more reliable results. Secondly, although subject-specific FE models of SCI, such as the ones developed in [3], have proved to be close representations of injury experiments, the acquired mechanical results are still approximations based on the defined geometries and available material properties. As these definitions are improved, the accuracy of the FE models will also increase, improving the datasets. The histology used to characterize tissue damage was obtained several weeks after the mechanical impact to the spinal cord, so the resulting damage is a combination of mechanical loading and biological responses, which likely confounds our ability to predict tissue level injury from mechanics alone. Additionally, there exist more sophisticated ML algorithms that might outperform the ones evaluated in this paper; however, this study was proposed as an exploratory attempt to validate and justify a more thorough exploration of the use of ML in the SCI context.

VI. Conclusion

This study explored one of the current limitations of the FE models of SCI: the identification of injured elements in the spinal cord. The results indicated that the performance of the ML models varied for GM, WM, and combined (GM & WM) datasets as the distribution of samples, relevant features, and target values varied across datasets. This suggests that FE models might benefit from using different classifiers to explore the correlation between the mechanical outputs and tissue damage in the gray and white matters. Combining the datasets from the GM and WM samples and defining the type of tissue as a feature showed little effect on improving the classification performance of the ML models. Moving forward, this approach can help define injury thresholds based on mechanical features and quantify relationships between mechanics and tissue damage. This will contribute to improving the pre-clinical reliability of computational models, and open new avenues for the implementation of ML algorithms to identify spinal cord injury damage.

References

[1] S. Mattucci, J. Speidel, J. Liu, B. K. Kwon, W. Tetzlaff, and T. R. Oxland, "Basic biomechanics of spinal cord injury — How injuries happen in people and how animal models have informed our understanding," *Clinical Biomechanics*, vol. 64, pp. 58–68, Apr. 2019, doi: 10.1016/j.clinbiomech.2018.03.020.

[2] C. J. Sparrey, E. A. Salegio, W. Camisa, H. Tam, M. S. Beattie, and J. C. Bresnahan, "Mechanical Design and Analysis of a Unilateral Cervical Spinal Cord Contusion Injury Model in Non-Human Primates," *Journal of Neurotrauma*, vol. 33, no. 12, pp. 1136–1149, Jun. 2016, doi: 10.1089/neu.2015.3974.

[3] S. Jannesar, E. A. Salegio, M. S. Beattie, J. C. Bresnahan, and C. J. Sparrey, "Correlating Tissue Mechanics and Spinal Cord Injury: Patient-Specific Finite Element Models of Unilateral Cervical Contusion Spinal Cord Injury in Non-Human Primates," *Journal of Neurotrauma*, vol. 38, no. 6, pp. 698–717, Mar. 2021, doi: 10.1089/neu.2019.6840.

[4] C. M. Russell, A. M. Choo, W. Tetzlaff, T.-E. Chung, and T. R. Oxland, "Maximum Principal Strain Correlates with Spinal Cord Tissue Damage in Contusion and Dislocation Injuries in the Rat Cervical Spine," *Journal of Neurotrauma*, vol. 29, no. 8, pp. 1574–1585, May 2012, doi: 10.1089/neu.2011.2225.

[5] J. T. Maikos, Z. Qian, D. Metaxas, and D. I. Shreiber, "Finite Element Analysis of Spinal Cord Injury in the Rat," *Journal of Neurotrauma*, vol. 25, no. 7, pp. 795–816, Jul. 2008, doi: 10.1089/neu.2007.0423.

[6] C. Persson, J. L. Summers, and R. M. Hall, "Modelling of Spinal Cord Biomechanics: In Vitro and Computational Approaches," in *Neural Tissue Biomechanics*, L. E. Bilston, Ed., in Studies in Mechanobiology, Tissue Engineering and Biomaterials, vol. 3. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 181–201. doi: 10.1007/8415_2010_38.

[7] O. Khan, J. H. Badhiwala, J. R. F. Wilson, F. Jiang, A. R. Martin, and M. G. Fehlings, "Predictive Modeling of Outcomes After Traumatic and Nontraumatic Spinal Cord Injury Using Machine Learning: Review of Current Progress and Future Directions," *Neurospine*, vol. 16, no. 4, pp. 678–685, Dec. 2019, doi: 10.14245/ns.1938390.195.

[8] A. Barrios-Anderson, J. S. Fridley, D. A. Borton, and C. Saab, "Decoding nociception in the spinal cord: Computer modeling and machine learning," in *Spinal Cord Injury Pain*, Elsevier, 2022, pp. 175–198. doi: 10.1016/B978-0-12-818662-6.00005-4.

[9] T. Inoue *et al.*, "XGBoost, a Machine Learning Method, Predicts Neurological Recovery in Patients with Cervical Spinal Cord Injury," *Neurotrauma Reports*, vol. 1, no. 1, pp. 8–16, Jan. 2020, doi: 10.1089/neur.2020.0009.

[10] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[11] E. A. Salegio *et al.*, "A Unilateral Cervical Spinal Cord Contusion Injury Model in Non-Human Primates (*Macaca mulatta*)," *Journal of Neurotrauma*, vol. 33, no. 5, pp. 439–459, Mar. 2016, doi: 10.1089/neu.2015.3956.

[12] C. Halimu, A. Kasem, and S. H. S. Newaz, "Empirical Comparison of Area under ROC curve (AUC) and Mathew Correlation Coefficient (MCC) for Evaluating Machine Learning Algorithms on Imbalanced Datasets for Binary Classification," in *Proceedings of the 3rd International Conference on Machine Learning and Soft Computing*, Da Lat Viet Nam: ACM, Jan. 2019, pp. 1–6. doi: 10.1145/3310986.3311023.

[13] R. J. Urbanowicz *et al.*, "A Rigorous Machine Learning Analysis Pipeline for Biomedical Binary Classification: Application in Pancreatic Cancer Nested Case-control Studies with Implications for Bias Assessments." arXiv, Sep. 08, 2020. Accessed: Feb. 11, 2023. [Online]. Available: http://arxiv.org/abs/2008.12829

[14] Q. Gu, L. Zhu, and Z. Cai, "Evaluation Measures of the Classification Performance of Imbalanced Data Sets," in *Computational Intelligence and Intelligent Systems*, Z. Cai, Z. Li, Z. Kang, and Y. Liu, Eds., in Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2009, pp. 461–471. doi: 10.1007/978-3-642-04962-0_53.

[15] S. Boughorbel, F. Jarray, and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric," *PLOS ONE*, vol. 12, no. 6, p. e0177678, Jun. 2017, doi: 10.1371/journal.pone.0177678.