Predictive Model Development to Identify Failed Healing in Patients after Non–Union Fracture Surgery

Cedric Donié*, Marie K. Reumann†, Tony Hartung†, Benedikt J. Braun†, Tina Histing†, Satoshi Endo*, Sandra Hirche*

Email: {cedric.donie,s.endo,hirche}@tum.de{mreumann,thartung,bbraun,thisting}@bgu-tuebingen.de

*Chair of Information-Oriented Control, Technical University of Munich, Munich, Germany

†Dept. of Trauma and Reconst. Surg., BG Klinik Tuebingen, Eberhard Karls University Tuebingen, Tuebingen, Germany

***Abstract*—Bone non-union is among the most severe complications associated with trauma surgery, occurring in 10–30 % of cases after long bone fractures. Treating non-unions requires a high level of surgical expertise and often involves multiple revision surgeries, sometimes even leading to amputation. Thus, more accurate prognosis is crucial for patient well-being.**

Recent advances in machine learning (ML) hold promise for developing models to predict non-union healing, even when working with smaller datasets, a commonly encountered challenge in clinical domains. To demonstrate the effectiveness of ML in identifying candidates at risk of failed non-union healing, we applied three ML models—logistic regression, support vector machine, and XGBoost—to the clinical dataset TRUFFLE, which includes 797 patients with long bone non-union.

The models provided prediction results with 70% sensitivity, and the specificities of 66 % (XGBoost), 49 % (support vector machine), and 43 % (logistic regression). These findings offer valuable clinical insights because they enable early identification of patients at risk of failed non-union healing after the initial surgical revision treatment protocol.

***Index Terms*—Machine learning, predictive models, non-union, bone healing, fracture healing, failed healing, pseudoarthrosis, personalized medicine**

I. Introduction

Bone non-union describes the failed healing of a fracture and represents one of the most severe complications encountered in the field of trauma surgery. The subsequent treatment necessitates complex surgical interventions, leading to a substantial reduction in patient quality of life [1]. Non-union treatment is challenging and frequently requires revision surgeries (Fig. 1). In some cases, revisions fail, making amputation the only option. Consequently, non-unions are a major socioeconomic burden [2]. Such treatment failures are partially attributable to the limited understanding of the factors that influence bone healing after non-union treatment. Predicting the outcome of the first non-union revision surgery would, therefore, be helpful to design more personalized interventions and increase the chance of recovery. Developing such a prediction model is challenging because the collection of non-union data is time-consuming and expensive, limiting the size of datasets.

This work has received funding from the European Research Council (ERC) Consolidator Grant Safe data-driven control for human-centric systems (CO-MAN) under Grant Agreement No. 864686.

The clinical need to predict whether patients will (continue to) suffer from a non-union is currently handled by the surgical expertise of individual specialists. Various heuristics have been developed to formalize this knowledge for the use in clinical practice. However, heuristic prediction methods have limited power due to their inability to capture complex patterns.

A novel approach is the use of machine learning (ML) to predict failed healing in patients with non-union after bone fracture. ML works well for risk and outcome prediction in other areas of medicine [3], and it has been used in prospective studies to predict whether non-union would occur after fractures. ML can predict the healing of vertebral fractures based on features such as medical imaging findings, body mass index (BMI), and age [4], [5] and for subtrochanteric femoral fractures based on age and treatment type [6]. However, this research only predicts non-union after very specific fractures. To our knowledge, no study has investigated predicting the healing of existing non-unions for different anatomical locations in long bones.

In this paper, we propose a model that successfully predicts failed healing after the first revision surgery for non-unions in various long bones. Our approach is investigated and validated with a dataset that was collected in a single trauma hospital.

II. Methods

A. Patient cohort

We collected data from 797 patients (67.9 % male, 32.1 % female sex, self-reported) with long bone non-union fractures at a level I trauma center from 1st January 2009 to 31st May 2023 (TRUFFLE database, ClinicalTrials.gov NCT06098157). Age ranged from 14 to 91 years (mean 50.1, SD 15.6). We received ethical approval from the University of Tuebingen ethical committee (840/2019BO2). A long bone non-union was defined as a fracture that had not shown any clinical and radiological signs of healing, with criteria adapted from the Radiographic Union Scale of Tibial Fractures (RUST) [7].

Fig. 1 illustrates the timeline of patients undergoing non-union treatment. During the follow-up visit after the first

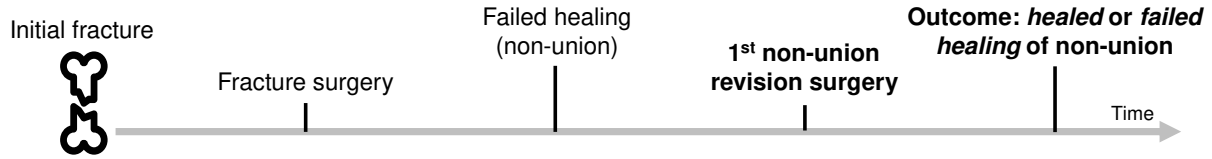

Fig. 1. Timeline of the clinical treatment of long bone non-unions

non-union revision surgery, orthopaedic and trauma specialists evaluated the treatment outcome: non-union healed or failed to heal. In the present dataset, failed healing was observed in 38.77 % of patients.

The dataset contained clinical features related to the initial trauma (i.e. date, anatomical location of fracture, severity of injury), to primary surgical treatment (i.e. date, type of surgical reduction and fixation), to screening at the time non-union diagnosis (i.e. age, sex, comorbidities, body mass index, Weber-Cech classification, biomechanical stability, soft tissue status, lab results, drug intake), and to non-union revision surgery (i.e. date, type of fixation, autologous bone grafting, use of growth factors, need of antibiotic treatment).

B. Data preprocessing

Classification requires features to be numerical for many approaches (e.g., in scikit-learn). We encode nominal values (categorical, e.g., type of osteosynthesis) with one-hot encoding, taking into account that more than one choice is possible for some features (e.g., comorbidities of diabetes and lung disease). This produces our dataset containing 356 features per patient with 16.6 % missing data. All dates (e.g., of the first fracture surgery or the first non-union revision surgery) are counted from the initial fracture date.

We split the data into 80 % training and 20 % test data. The proportion of patients with failed healing is equivalent for training (38.77 %) and test (38.75 %) data. To avoid overfitting, we hold out the test data, reserving 20 % of training data for intermediate model evaluation instead.

C. Model development

XGBoost, support vector machines (SVMs), and logistic regression were analyzed in the present work. Each of these models outputs a predicted probability of failed non-union healing. To convert the probability to a binary prediction (not healed vs. healed), a threshold was chosen and all probabilities above this threshold are were rated as failed healing. Unless stated otherwise, we used a threshold of 0.5.

XGBoost is an ensemble of *regression trees* [8]. Regression trees are similar to decision trees, but output an estimated likelihood rather than a binary prediction. XGBoost trains many regression trees on subsets of the training data via *boosting* and averages their output to yield a class likelihood. Boosting means that new trees in the ensemble depend on previously trained trees. We selected XGBoost for further study because it is state-of-the-art for tabular data classification [8], [9], can model non-linear relationships well [3], and was used in related non-union work [4], [5].

SVMs classify high-dimensional input into two classes by treating each feature as a dimension and linearly separating the features with a *hyperplane*. For example, two features would be separated by a line (1d hyperplane), three features by a plane (2d hyperplane), and n features by a $(n-1)$-dimensional hyperplane [10]. Non-linear classification is possible by transforming the feature space with a radial base function. Probabilities are obtained by applying a logistic function to the SVM's output [11]. SVM was chosen because it is common in contemporary research [3] and it has seen time-testing than XGBoost.

Logistic regression outputs the probability of the positive class via the logistic function applied to a linear combination of all the features [12]. An optimization algorithm selects regression coefficients that minimize the prediction error. We chose logistic regression as a baseline because it is simple, the most common medical prediction model, and works comparatively well for the oftentimes small clinical datasets encountered in clinical settings [12].

For XGBoost, the positive class is weighted with the ratio of healed to non-healed patients in the training data, compensating for the slight class imbalance. To reduce the chance of overfitting that would reduce the performance on external validation, we opted for the standard XGBoost hyperparameters, which are known to work on a wide variety of datasets [9]. However, we limited the tree depth to five, as performance deteriorated with deeper trees in our initial experiments.

Logistic regression and SVM benefit from scaled input and cannot handle missing values. Thus, standard scaling is applied to all ordinal (e.g., Weber-Czech [13] classification into hypertrophic, oligotrophic, and atrophic), interval (e.g., number of previous surgeries), and continuous (e.g., hemoglobin) values. Standard scaling z of a single value x is defined as $z = (x - \mu)/\sigma$, where μ is the mean and σ is the standard deviation over all patients. Missing numerical values are replaced with the feature's mean and missing booleans/categoricals with the feature's most frequent value.

D. Model evaluation

For the evaluation of a classifier, the first step is to calculate a single metric summarizing the classifier's performance. The next step is to compared the performance metrics between classifiers and consider statistical significance. In addition to the discriminative ability of the model, it is critical to evaluate model calibration to understand how well the predicted risk of non-union matches the actual incidence of non-union.

1) Performance metrics: The unified performance measure (UPM) [14] and the Matthews correlation coefficient

(MCC) [15] are most frequently recommended [15]. We opted for the UPM, which can be calculated directly from the confusion matrix according to [14].

$$\begin{aligned}\text{UPM} &= \frac{4 \cdot TP \cdot TN}{4 \cdot TP \cdot TN + (TP + TN) \cdot (\text{FP} + \text{FN})} \\ &= \frac{4}{\frac{1}{\text{Precision}} + \frac{1}{\text{Sens.}} + \frac{1}{\text{Spec.}} + \frac{1}{\text{Neg. Predictive Value}}}\end{aligned}$$

We examined multiple thresholds to compensate for the fact that UPM depends on a single confusion matrix (i.e., threshold).

2) Statistics for model comparison: It is important to ensure that differences between several classifiers' UPM scores are statistically significant, requiring multiple trainings and evaluations of the same classifier. However, all classifiers are deterministic during training, which means that repeated training of the same model would yield identical scores. To provide a random component, we generated new training datasets by randomly sampling 80 % of the full training dataset without replacement. This was repeated 300 times to generate 300 unique training datasets. We trained XGBoost, logistic regression, and SVM for each of these 300 datasets. The Wilcoxon signed-rank test compared the performance of the classifiers pairwise. The effect size r was estimated as $r = Z/\sqrt{N}$, where Z is the normalized test statistic and N is the number of samples. We set the significance level at $\alpha = 0.05$ and used Bonferroni correction for multiple comparisons.

3) Model calibration: To measure how well calibrated a model is, we regressed each patient's actual healing/failed healing against the predicted probability of failed healing by means of locally weighted scatterplot smoothing (LOWESS) (with the *statsmodels* implementation defaults of 2/3 of the total data for each y and three residual-based re-weightings), as suggested by [12]. The regression was applied to all of the test data. Furthermore, the mean bias of our model was expressed as an odds ratio (OR) of predicted incidence $\hat{y}$ vs. actual incidence y.

$$\begin{aligned}\text{OR} &= \text{odds}(\text{mean}(\hat{y}))/\,\text{odds}\,(\text{mean}\,(y)) \\ &= \frac{\text{mean}(\hat{y})/\,(1 - \text{mean}(\hat{y}))}{\text{mean}\,(y)\,/\,(1 - \text{mean}\,(y))}\end{aligned}$$

E. Ablation studies

To understand how the results would change with the amount of patient data, we artificially reduced the training data. Specifically, we defined multiple fractions of the training data from zero to one. For each of these fractions, we generated 25 samples (with replacement) from the entire training dataset, trained XGBoost, and evaluated the performance (UPM, sens., spec.) over the whole test dataset.

III. Results and Discussion

All three models delivered a good predictive performance with 70 % sensitivity and specificities of 66 % for XGBoost, 49 % for SVM, and 43 % for logistic regression.

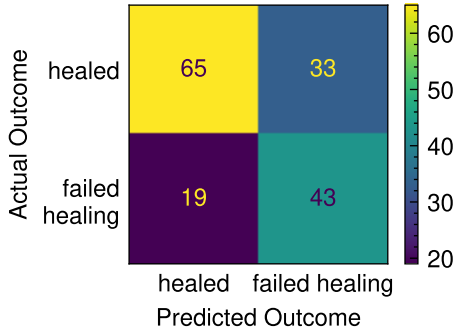

Fig. 2. Confusion matrix for XGBoost with a threshold of 0.26. This threshold was chosen to guarantee at least 70 % sensitivity.

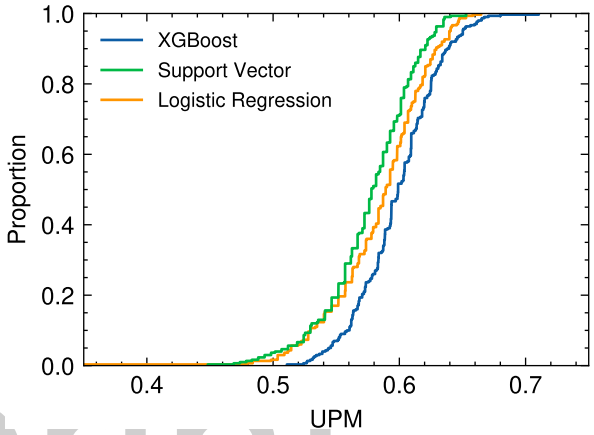

Fig. 3. Empirical cumulative distribution functions of XGBoost, SVM, and logistic regression. Each classifier is trained 300 times on randomly sampled 80% of the training data. A lower curve indicates stochastic dominance.

In terms of UPM, XGBoost outperforms the other classifiers with statistical significance. Fig. 3 shows that XGBoost is generally stochastically dominant over SVMs and logistic regression ($\alpha = 0.05$). XGBoost UPM was persistently superior and robust across the threshold range (Fig. 4). To provide a concrete example, Fig. 2 shows the confusion matrix for XGBoost with a threshold of 0.26. This threshold was chosen to obtain at least 70 % specificity. XGBoost was well-calibrated in the sense that a higher predicted probability generally corresponded to a higher risk of failed non-union healing, (Fig. 5). However, XGBoost does not exactly estimate the probability of failed healing. The prediction has a calibration odds ratio of 0.823 vs. the actual risk of failed non-union healing. Due to this imperfect calibration, the XGBoost confidence cannot be used to estimate the true healing likelihood (i.e., aleatoric uncertainty).

Finally, Fig. 6 shows that artificially decreasing the number of patients in the training dataset only slightly decreases the performance, as measured by UPM. This further underlines that our XGBoost model performs well with real-world clinical datasets which are limited in size.

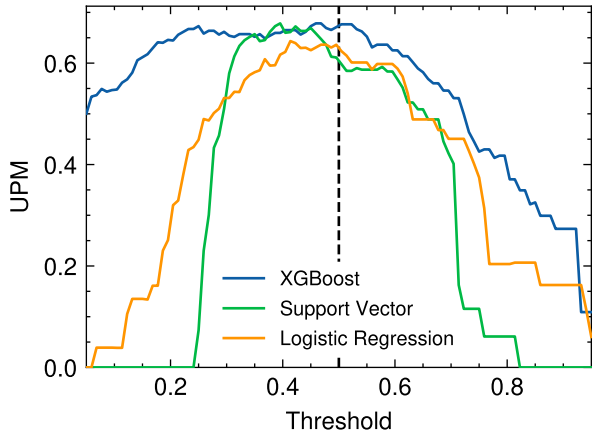

Fig. 4. UPM is only slightly affected by the chosen decision threshold. UPM is calculated with different thresholds above which a prediction is rated positive.

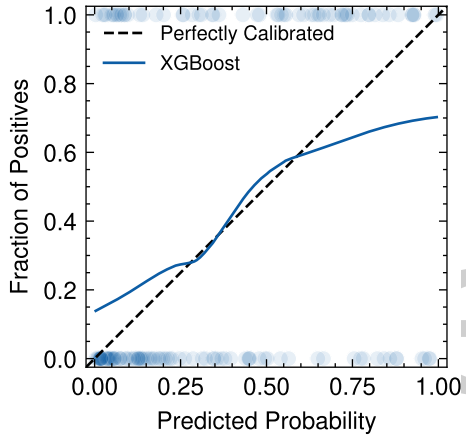

Fig. 5. Calibration display for XGBoost based on the test data. The true class is shown in the scatter plot. This scatter plot is smoothed using LOWESS.

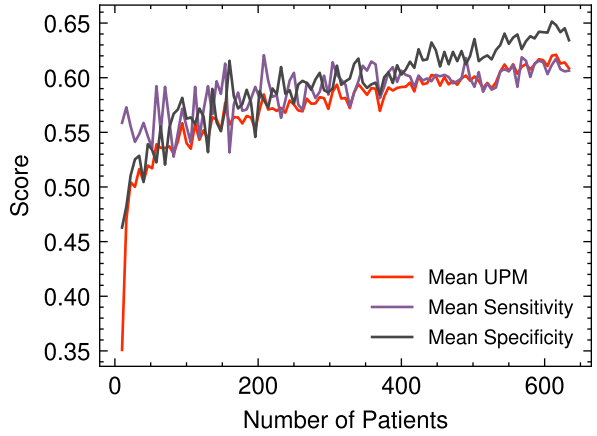

Fig. 6. UPM, sensitivity, and specificity of XGBoost increase with more patients in the training dataset. Each data point is generated by sampling the entire training dataset 25 times, training the model, and calculating metrics on the test dataset. The threshold is chosen at 0.26.

IV. Conclusion

ML models, especially XGBoost, are suited to identifying patients at risk of failed healing after non-union revision surgery. Clinically relevant predictive performance has been demonstrated using data from a single center. Future research should consider investigating the performance with larger, multi-center datasets. Our results pave the way to identifying patients at risk of failed non-union healing and potentially allow for more personalized treatment of the debilitating non-unions encountered in trauma surgery.

Acknowledgment

We thank Dr. Dr. Matthias Reumann for his feedback throughout the research process.

References

[1] M. R. Brinker, C. M. Loftis, J. D. Khoriaty, and W. R. Dunn, "The devastating effects of humeral nonunion on health-related quality of life," *J. Shoulder Elbow Surgery*, vol. 31, pp. 2578–2585, Dec. 2022.

[2] D. Saul, M. M. Menger, S. Ehnert, A. K. Nüssler, T. Histing, and M. W. Laschke, "Bone healing gone wrong: Pathological fracture healing and non-unions—overview of basic and clinical aspects and systematic review of risk factors," *Bioengineering*, vol. 10, p. 85, Jan. 2023.

[3] O. Elfanagely, Y. Toyoda, S. Othman, J. A. Mellia, M. Basta, T. Liu, K. Kording, L. Ungar, and J. P. Fischer, "Machine learning and surgical outcomes prediction: A systematic review," *J. Surgical Res.*, vol. 264, pp. 346–361, Aug. 2021.

[4] S. Takahashi, H. Terai, M. Hoshino, T. Tsujio, M. Kato, H. Toyoda, A. Suzuki, K. Tamai, A. Yabu, and H. Nakamura, "Machine-learning-based approach for nonunion prediction following osteoporotic vertebral fractures," *Eur. Spine J.*, Oct. 2022.

[5] I. Leister, T. Haider, M. Vogel, J. Vastmans, P. Langthaler, G. Mattiassich, A. Christ, M. Etschmaier, A. Eijkenboom, J. Burghuber, H. Kindermann, O. Mach, D. Maier, and F. Högel, "A predictive model to identify treatment-related risk factors for odontoid fracture nonunion using machine learning," *Spine*, vol. 48, p. 164, Feb. 2023.

[6] P. Nag and S. Chanda, "A preclinical model of post-surgery secondary bone healing for subtrochanteric femoral fracture based on fuzzy interpretations," *PLOS ONE*, vol. 17, p. e0271061, July 2022.

[7] D. B. Whelan, M. Bhandari, D. Stephen, H. Kreder, M. D. McKee, R. Zdero, and E. H. Schemitsch, "Development of the radiographic union score for tibial fractures for the assessment of tibial fracture healing after intramedullary fixation," *J. Trauma Acute Care Surgery*, vol. 68, p. 629, Mar. 2010.

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, KDD'16, (New York, NY, USA), pp. 785–794, Association for Computing Machinery, Aug. 2016.

[9] V. Borisov, T. Leemann, K. Sessler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep neural networks and tabular data: A survey," *IEEE Trans. Neural Netw. Learning Syst.*, pp. 1–21, Dec. 2022.

[10] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273–297, Sept. 1995.

[11] J. C. Platt, "Probabilities for SV machines," in *Advances in Large Margin Classifiers*, no. 11 in NIPS, pp. 61–75, MIT Press, Oct. 2023.

[12] E. W. Steyerberg, *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating*. Statistics for Biology and Health, Cham: Springer International Publishing, 2019.

[13] B. G. Weber and O. Čech, *Pseudarthrosis: Pathophysiology, Biomechanics, Therapy, Results*. Bern: Hans Huber Publishers, 1976.

[14] A. R. Redondo, J. Navarro, R. R. Fernández, I. M. de Diego, J. M. Moguerza, and J. J. Fernández-Muñoz, "Unified performance measure for binary classification problems," in *IDEAL 2020* (C. Analide, P. Novais, D. Camacho, and H. Yin, eds.), Lecture Notes in Computer Science, (Cham), pp. 104–112, Springer International Publishing, 2020.

[15] D. Chicco and G. Jurman, "The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification," *BioData Mining*, vol. 16, p. 4, Feb. 2023.