Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online
CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 48-20 (2015) 469–474

Machine Learning for Predictive Modelling based on Small Data in Biomedical Engineering

Torgyn Shaikhina[1], Dave Lowe[2], Sunil Daga[3, 4], David Briggs[3], Robert Higgins[4] and Natasha Khovanova[1]

[1]School of Engineering, University of Warwick, Coventry, CV47AL UK
(tel.: +44(0)2476528242; e-mail: N.Khovanova@warwick.ac.uk)
[2] NHS Blood and Transplant Birmingham [3]Warwick Medical School
[4]University Hospitals Coventry and Warwickshire NHS Trust

Abstract: Experimental datasets in bioengineering are commonly limited in size, thus rendering Machine Learning (ML) impractical for predictive modelling. Novel techniques of multiple runs for model development and surrogate data analysis for model validation are suggested for prediction of biomedical outcomes based on small datasets for classification and regression tasks. The proposed framework was applied to designing a Neural Network model for osteoarthritic bone fracture risk stratification, and a Decision Tree model for prediction of antibody-mediated kidney transplant rejection. Despite the small datasets (35 bone specimens and 80 kidney transplants), the two models achieved high accuracy of 98.3% and 85%, respectively.

Keywords: Machine Learning, Small Data, Biomedical Systems, Decision Tree, Neural Network

1. INTRODUCTION

Machine Learning (ML) enables data-driven models to "learn" information about a system directly from observed data without predetermining mechanistic relationships that govern the system. Due to the ability of ML algorithms to adaptively improve their performance with each new data sample, ML has become the core technology for numerous real-world applications: from weather forecasting and DNA sequencing, to Internet search engines and stock market predictions. Nevertheless, ML systems are rarely viewed in the context of small data, where insufficient number of training samples can compromise the learning success (Forman & Cohen 2004; Lanouette et al. 1999).

Small dataset conditions (less than 10 occurrences per predictor variable) are characteristic of biomedical engineering domain, where complexity and the high cost of experiments restrain the number of available samples (Hudson & Cohen 2000). It has been argued that ML can offer an indispensable tool for biomedical problems involving complex heterogeneous data when conventional statistical tools fail (Inza et al. 2010; Campbell 2014; Grossi 2011). In applications such as gene selection (Hoff et al. 2008), screening heart murmurs in children (DeGroff et al. 2001), and predicting breast cancer relapse (Faradmal et al. 2014), ML-based models were able to map highly non-linear input and output patterns even when mechanistic relationships between model variables could not be determined due to pathologies or complexity.

Nonetheless, the vast potential of ML for predictive modelling in bioengineering remains largely unexplored. To extend the benefits of ML to a wider range of bioengineering models, it is essential to develop methods that would cope with the limited data size.

This work was supported by the EPSRC UK Grant EP/K02504X/1.

In their early work, the authors have been successful in applying Neural Networks (NNs) to a correlation analysis of a small dataset of 35 samples in hard tissue engineering (Khovanova et al. 2014; Shaikhina et al. 2014). Current paper develops a ML framework for predictive modelling based on small biomedical datasets for classification and regression tasks. Specifically, the paper considers two biomedical applications:

a. NNs for prediction of Compressive Strength (CS) of human trabecular bone in severe osteoarthritis (regression model).

b. Decision Trees (DTs) for prediction of acute antibody-mediated rejection (ABMR) of kidney transplants based on pre-operative clinical indicators (classification model).

2. METHODOLOGY

2.1 Comparing ML model designs using multiple runs

ML algorithms commonly contain a deliberate degree of randomness in their training and initialization routines. Random starting points are often necessary to improve the algorithm's convergence to the global minimum (Forman & Cohen 2004; Hudson & Cohen 2000). This comes with a negative side effect on the algorithm stability and generalisation, which becomes more pronounced when only a small number of training samples is available. In other words, a ML algorithm trained on a small dataset may produce dissimilar output patterns depending on the random initial conditions. Subsequently, various instances of the same small-data ML model would often exhibit erratic fluctuations in performance. This prevents effective comparison between ML models and hinders the possibility of their optimisation.

We introduce a method of multiple runs in order to provide means for consistent comparisons between various ML models, which enables their subsequent optimisation. First, for

2405-8963 © 2015, IFAC (International Federation of Automatic Control) Hosting by Elsevier Ltd. All rights reserved.
Peer review under responsibility of International Federation of Automatic Control.
10.1016/j.ifacol.2015.10.185

a given ML model a large number of instances with various initial conditions are generated and trained in parallel. Consequently the performance of the ML model is assessed not on a single instance, but repeatedly on a set of a few thousands of instances of the same model (hereafter *run*). The optimal design is then determined by comparing the average performance between runs of various ML models, even when individual instances cannot be compared. Once optimal model design is selected, the single best performing instance of that design is used as the final model.

When applied to NNs and DTs, this strategy principally differs from the ensemble-NNs or Random Forests in that only the output of the best performing NN/DT instance is ultimately selected as the final predictive model.

Choice for the size of the run, i.e. how many model NN/DT instances each run contains, was influenced by the need to balance between desired precision of performance measures and computational efficiency, as larger runs require more memory and time to simulate. For the two applications in this study we found that the minimum sizes of the run that maintained performance measures consistent to 3 decimal places were 2000 for NNs and 600 for DTs.

2.2 Validating ML models for regression tasks using surrogate data

Small dataset conditions and the associated random effects make validation of ML models for regression tasks impractical. Conventional methods, such as cross-validation, may become unreliable when the number of independent test samples is limited. This necessitates an alternative approach for validating regression ML models in the presence of random effects due to small data.

Inspired by the success of the surrogate data approach (Theiler et al. 1992; Hirata et al. 2008) for biomedical and nonlinear physics applications, and neural coding, we propose to use surrogate data for validation of regression ML models built on small data. The surrogates were generated from random numbers to mimic the distribution of the original dataset independently for each component of the input vector. While resembling the original data statistically in terms of their mean, standard deviation and range, the surrogates do not retain the intricate interrelationships between the variables of the real dataset. Hence successful real-data models are expected to perform significantly better than the surrogate data models.

In our proposed framework, validation with surrogate data was considered in the context of multiple runs and was used for comparison of the real-data NN model of the optimal design with the surrogate-data NN of the same design on a run of 2000 NN instances. To improve robustness, the experiment was replicated in 10 runs involving 20000 NNs in total.

We demonstrate on the example of NNs that a ML regression model trained and tested on surrogates can be used as a benchmark for validating real data models by setting a performance threshold for the random effects due to small data. Defined as the highest performance achieved by surrogate models, this threshold indicates the lower performance boundary expected of the real data models.

3. DATA MODELS

3.1. Regression NN for osteoarthritic bone CS prediction

The NN model was designed to predict the CS of an osteoarthritic trabecular bone from micro–CT indications of its morphology, level of interconnectivity, porosity, as well as patient's gender and age. The detailed description of the dataset, comprising 35 human femora, can be found in the original study by Perilli et al. (2007). The samples were divided into training (22 samples) and validation (6 samples) sets using a random permutation, while remaining samples were reserved for test (7 samples) and fixed for every NN.

Considering the size and the nature of the available data, a two-layer feedforward backpropagation NN was chosen as the base for the CS model with 5 input features and 1 output (Fig.1). The heterogeneous 1x5 input vector, $\bar{x}$, was stacked in the following order: x_1 = Structure Model Index (SMI), x_2 = trabecular thickness (Tb.Th), x_3 = bone volume density (BV/TV), x_4 = age and x_5 = gender. The 5x4 input weights matrix, IW, 4x1 layer weights column vector, $\overline{lw'}$, and the corresponding biases $\overline{b_{(1)}}$ and $b_{(2)}$ for each layer were initialized according to the Nguyen-Widrow method (Nguyen & Widrow 1990) in order to distribute the active region of each neuron in the layer approximately evenly across the layer's input space. Neurons in the hidden layer implemented a hyperbolic tangent sigmoid transfer function (Yonaba et al. 2010), while the output neuron computed the CS output from the input using a simple linear transfer function.

NNs were trained using Leverberg-Marquardt back-propagation algorithm (More 1978). The cost function was evaluated by the mean squared error between the output and actual CS values. The NN performance was measured by regression factor, R, between the actual CS values and the values predicted by the NN. The techniques of early-stopping and cross-validation were implemented in order to avoid NN overtraining and hence ensured better generalisaiton (Fushiki 2009). The resulting NN model mapped the output, y (in MPa) to the input vector, x:

$$y = \tanh\left[\bar{x} \cdot IW + \overline{b_{(1)}}\right] \cdot \overline{lw'} + b_{(2)} \quad (1)$$

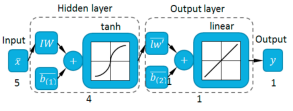

Fig.1. NN model topology and layer configuration is represented by a 5D input vector, a hidden layer with 4 neurons and a single output neuron.

The method of multiple runs was used to determine the optimal NN topology in terms of the hidden layer size and the early-stopping cut-off. Hidden layer sizes from 1 to 13 neurons and the early-stopping cut-off factors from 1 to 10 have been considered. Each of these 130 topology configurations was evaluated in a run of 2000 NNs. The optimal NN configuration

(Fig.1) comprised 4 neurons in the hidden layer and utilised a cut-off of 9 for early-stopping. Such iterative process using multiple runs involved 260000 individual NNs in total, but only the best performing NN from the 2000 with the optimal topology configuration, was selected as the final model for predicting CS in this application.

3.2. Classification DT for prediction of early kidney rejection

A classification model was developed for the prediction of acute/early antibody-mediated rejection within first 30 days after operation. The clinical focus of the model was to investigate how pre-treatment donor specific antibody (DSA) Immunoglobulin G (IgG) subclass levels, measured using Median Fluorescence Intensity (MFI) cytometry techniques, affect early outcome of transplantation when accounted for multiple baseline characteristics (Lowe et al. 2013). The following 15 parameters, measured before operation, were considered as input variables for the predictive model:

- 7 *continuous:* highest IgG DSA MFI level, patient's age, years on dialysis, and 4 total IgG subclass MFI levels,
- 4 *categorical*: cytometery cross-match (bead, flow or CDC), total number of HLA mismatches between donor and recipient (0-6), the number of class II HLA-DR mismatches (0-2), and the number of previous transplants (0-2),
- 4 *binary:* gender (male/female), delayed graft function (yes/no), live/deceased donor, and the presence of ClassI and ClassII (yes/no) HLA DSA.

The output ABMR was a two-class binary variable, where ABMR+ve corresponded to the early rejection of kidney (class 1), and ABMR-ve corresponded to an outcome without early rejection (class 0).

The dataset featured 80 samples: 60 observations were used for model training and the remaining 20 samples were reserved for tests. The data was well balanced (46 ABMR+ve and 34ABMR-ve samples), but contained 3 samples with missing data, which were excluded from the training pool.

Among multiple ML classifiers, DTs are particularly well suited for this task as they can perform automatic feature selection and thus are able to reduce the dimensionality of the data (Podgorelec et al. 2002; Azar & El-Metwally 2013). As implied in its name, DT is a tree-like structure, where leaf nodes represent class labels and branch nodes represent conjunctions of input features that resulted to those class labels.

The DT design in the present study was based on the standard CART algorithm implemented using MATLAB™ (Breiman et al. 1984). Throughout the training process, the dataset was recursively divided according to the split criterion until the optimal DT hierarchy of nodes was reached. The split optimisation criterion used in this DT model is the Gini's Diversity Index (GDI), which is a measure of node impurity. The node is considered pure when it contains only observations of one class (either ABMR+ve or ABMR-ve); the GDI of a pure node is equal to 0 (Coppersmith et al. 1999). The following additional constraints were imposed on the DT size: minimum 10 observations for the node to become a branch node and at least 1 observation per a leaf node.

Notably, for a DT classifier, finding an optimal binary split for a continuous predictor is far less computationally intensive than for a categorical predictor with multiple levels. In former case, DT can split between any two adjacent values of a continuous vector, but for a categorical predictor with i levels, all of the 2^{i-1}–1 splits need to be considered to find the optimal one. As an example: to identify the optimal split for the total number of HLA mismatches (i=7) the DT had to consider 127 possibilities.

To avoid overfitting, 10-fold cross-validation was implemented in the DT design (Fushiki 2009). Using the method of multiple runs presented in section 2.1, 600 individual DTs were generated and the best performing model was selected.

4. RESULTS

4.1. NN model

The optimal NN predicted CS with a root-mean-square error (rmse) of 0.85 MPa. The linear regression factors, R, between the actual and predicted CS, were 99.9% across the entire dataset and 98.3% on tests (Table I).

The surrogate data approach described in section 2.2 revealed significant differences in performance between real and surrogate NNs (Fig.2). Evaluated across 10 runs of 2000 NN instances, the real data NNs consistently outperformed surrogate NNs with the corresponding increase in $\mu(R_{all})$ from 0.33 to 0.68 (Fig.2). The Wilcoxon rank sum test confirmed the statistical difference between the medians of the regression factors achieved by the two models with p<0.000001. The median $R_{sur,all}$ was 0.38 for the surrogate NNs versus median $R_{real,all}$ of 0.78 for the real data NNs. The surrogate threshold defined in section 2.2 as the highest-performing surrogate model, was equal to $max(R_{sur,all})$ = 0.87. This threshold was exceeded by optimal real-data NN model, thus validating its performance in the presence of random effects due to small data.

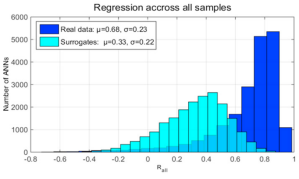

Fig.2. Comparison of performance regression coefficients, R_{all}, between NNs trained on surrogate data (light blue) and real bone samples (dark blue). The mean, μ, and the standard deviation, σ, were calculated across 20000 NNs.

4.2. DT model

The optimal DT achieved the classification accuracy, C = 86.7% during the training phase and correctly classified 85% of test cases (Table I). The DT identified ABMR+ve patients with 88.9% sensitivity and ABMR-ve cases with 82.9% specificity. The area under the curve (AUC) of the receiver operating characteristic (ROC) was equal 0.852 for the DT predictions on tests (Fig.3).

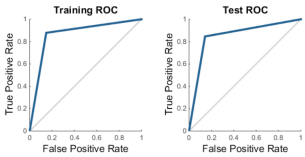

Fig.3. ROC curves for DT classification accuracy on the training dataset (left) and generalising performance on the test samples (right).

Out of 15 possible predictors, DT identified the following 6 variables as key to ABMR prediction (top to bottom in Fig. 4): the highest MFI DSA level (igg_hi), total IgG4 MFI (igg4), HLA mismatch number (mm), total IgG2 MFI (igg2), delayed graft function (dgf) and the total IgG1 MFI (igg1).

Additionally, the node splits in the DT hierarchy (Fig.4) provide an indication as to what specific *levels* of the HLA DSA antibodies were statistically associated with each of the ABMR+ve/ABMR-ve classes. For instance, the DT identified that patients with the highest MFI DSA levels below 834 belong to the ABMR-ve group, while those with igg_hi ≥ 834 and igg4 ≥ 36.5 have a high likelihood of early transplant rejection. Similarly, patients with 4 or 5 HLA mismatches belong to the ABMR+ve group.

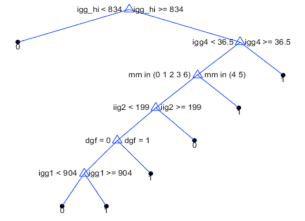

Fig.4. DT model schematic showing the split hierarchy based on 6 branch nodes and 7 leaf nodes.

Table I – Predictive performance of the two ML models

Model	Performance measure	performance on data subsets			
		training	*validation*	*test*	*all*
NN	regression, *R*	0.999	0.991	0.983	0.993
DT	classification, *C*	0.867	n/a	0.850	0.862

5. DISCUSSION

5.1. Practical significance of the NN-based osteoarthritic bone model

Compressive strength of a trabecular tissue in femoral head is indicative of the bone fracture risk, which accounts for more than 20% of orthopaedic hospital cases in the UK and its global incidence is projected to affect 6.26 million people by 2050 (Johnel et al. 1992; Dhanwal et al. 2011). Computational techniques such as quantitative computer tomography-based finite element analysis (FEA) and the method of indirect estimation through densitometry have been commonly used for clinical data in hip fractures (Geris 2013; Bessho et al. 2004; Keyak et al. 1997), yet case-specific models in patients affected by degenerative bone diseases, such as osteoarthritis, are scarce (Tseng et al. 2013).

Furthermore, the established mathematical relationships between the structural parameters and the mechanical strength in healthy bones does not hold for the damaged trabecular tissue, making the mechanistic modelling impractical (Helgason et al. 2008). To the authors' best knowledge, the NN presented in this study is the only existing patient-specific predictive model for trabecular bone CS in severe osteoarthritis.

The NN model relates heterogeneous biological and structural parameters to output specific values of the specimen's CS. Despite the small number of samples available for the model development, the NN was successful in mapping the 5-dimensional input to the continuous CS vector with an accuracy of 99.3%. The comparably high accuracy was achieved on test samples with which NN had not been previously presented: the regression coefficient between the actual and predicted CS values evaluated on the test samples was equal to 98.3%. This is 8.7% more accurate than the existing mechanistic model based on bivariate power regression between CS and bone porosity (Perilli et al. 2007).

Notably, all of the 5 biological and structural indicators used as the NN model inputs could be determined from computer tomography scans, thus allowing for a non-destructive estimation of trabecular tissue fracture risk. This accurate, patient data driven model for CS estimation based on NN can be used by hard tissue engineers when designing bioscaffolds that imitate the natural trabecular bone (Geris 2013) and offer a clinical decision support tool for diagnosis, prevention and potential treatment of osteoarthritis (Sinusas 2012; Amato et al. 2013).

5.2. Practical significance of using DT for predictive modelling of ABMR in renal transplantation

The classification accuracy of 85% achieved by the DT model on independent test samples demonstrates that the ML approach can be effectively applied for predictive modelling in renal transplantation despite the small number of observations and heterogeneous input parameters. The ROC curves (AUC = 0.85) of the proposed DT model are not dissimilar to the clinical prognosis by a human expert (Podgorelec et al. 2002). In comparison, a DT-based model for long-term kidney allograft survival developed by Krikov et al., achieved AUC of 0.63, 0.64, 0.71, 0.82, and 0.90 for the 1, 3, 4, 5, and 10 year outcome predictions, respectively (2007). Krikov's model was built on data from 92,844 patient records from the US Renal Data System, versus 80 samples available for our DT model with an equally high performance for short-term rejection.

The DT model was able to successfully determine the optimal hierarchy of parameters associated with early kidney rejection. The 6 key predictors identified by the DT were confirmed by Logistic Regression Likelihood model, which is a tool of choice in medical statistics for binary classification (Zhu et al. 2013; Behera et al. 2011; Bouwmeester et al. 2012). However, the superiority of the DT model is that it was also able to determine dangerous antibody levels, which conventional statistical methods were not able to provide.

Predictions made by our DT are patient-specific and are based on clinical indicators either known prior to the surgery or in the immediate post-transplant period, thus yielding a tool for accurate ABMR risk stratification preceding individual transplantation.

5.3. The significance of the generalised framework for small dataset conditions in bioengineering

The two successful applications of the proposed multiple run technique for both classification and regression tasks confirm that the size of the available dataset does not necessarily limit the utility of ML-based methods in the biomedical domain. The accuracies of the NN and DT models trained, validated and tested on as few as 35 samples, were comparable to some of the highest performing ML models designed on significantly larger datasets (200-9000 samples) in related biomedical applications, ranging from the predictions of hip fractures in osteoporosis (Hu et al. 2012), genotype- phenotype risk patterns in diabetic kidney disease (Leung et al. 2013), 2-hour plasma glucose of the 75g Oral Glucose Tolerance Test (Gao et al. 2011) to the above-mentioned model for long term kidney graft survival (Krikov et al. 2007). Hence, despite the small data conditions, the proposed framework allows for ML techniques to be successfully used for predictive modelling in bioengineering.

6. CONCLUSIONS

1) Sporadic effects of ML performance due to insufficient samples can be mitigated by evaluating a large set of ML models (≥2000). The proposed approach of multiple runs allowed for consistent comparisons between various ML configurations.

2) ML models for *regression* tasks based on small data can be successfully validated using synthesised 'surrogates'. A large-scale study involving 20000 NNs confirmed that models trained on real biomedical data achieve a 35% increase in the average predictive performance than analogous models trained on the surrogate controls.

3) Our framework allowed ML algorithms to successfully predict biomedical outcomes despite small datasets in both classification and regression tasks. The NN regressive model achieved 98.3% accurate predictions of CS in trabecular bones affected by osteoarthritis, while the DT classifier was 85% accurate in determining the ABMR of a kidney transplant from baseline clinical characteristics, including pre-transplant antibody levels.

REFERENCES

Amato, F. et al., 2013. Artificial neural networks in medical diagnosis. *Journal of Applied Biomedicine*, 11, pp.47–58.

Azar, A. & El-Metwally, S., 2013. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7-8), pp.2387–2403.

Behera, M. et al., 2011. Statistical learning methods as a preprocessing step for survival analysis: evaluation of concept using lung cancer data. *BioMedical Engineering OnLine*, 10(1), pp.10–97.

Bessho, M. et al., 2004. Prediction of the strength and fracture location of the femoral neck by CT-based finite-element method: A preliminary study on patients with hip fracture. *Journal of Orthopaedic Science*, 9, pp.545–550.

Bouwmeester, W. et al., 2012. Reporting and Methods in Clinical Prediction Research: A Systematic Review. *PLoS Med*, 9(5), p.e1001221.

Breiman, L. et al., 1984. *Classification and Regression Trees*, New York: Chapman & Hall.

Campbell, C., 2014. Machine Learning Methodology in Bioinformatics. In N. Kasabov, ed. *Springer Handbook of Bio-/Neuroinformatics*. Springer Berlin Heidelberg, pp. 185–206.

Coppersmith, D., Hong, S.J. & J., H., 1999. Partitioning Nominal Attributes in Decision Trees. *Journal of Data Mining and Knowledge Discovery*, 3, pp.197–217.

DeGroff, C.G. et al., 2001. Artificial neural network-based method of screening heart murmurs in children. *Circulation*, 103, pp.2711–2716.

Dhanwal, D.K. et al., 2011. Epidemiology of hip fracture: Worldwide geographic variation. *Indian Journal of Orthopaedics*, 45(1), pp.15–22.

Faradmal, J. et al., 2014. Comparison of the performance of log-logistic regression and artificial neural networks for predicting breast cancer relapse. *Asian Pacific journal of cancer prevention : APJCP*, 15(14), pp.5883–5888.

Forman, G. & Cohen, I., 2004. Learning from Little: Comparison of Classifiers Given Little Training. *Proc PKDD*, 19, pp.161–172.

Fushiki, T., 2009. Estimation of prediction error by using K-fold cross-validation. *Statistics and Computing*, 21, pp.137–146.

Gao, W. et al., 2011. Using neural network as a screening and educational tool for abnormal glucose tolerance in the community. *Archives of Public Health*, 68(4), pp.143–154.

Geris, L., 2013. *Computational Modeling in Tissue Engineering*, Berlin: Springer-Verlag.

Grossi, E., 2011. Artificial Neural Networks and Predictive Medicine: a Revolutionary Paradigm Shift. In K. Suzuki, ed. *Artificial Neural Networks - Methodological Advances and Biomedical Applications*. InTech, pp. 139–150.

Helgason, B. et al., 2008. Mathematical relationships between bone density and mechanical properties: A literature review. *Clinical Biomechanics*, 23, pp.135–146.

Hirata, Y. et al., 2008. Testing a neural coding hypothesis using surrogate data. *Journal of Neuroscience Methods*, 172, pp.312–322.

Hoff, K. et al., 2008. Gene prediction in metagenomic fragments: A large scale machine learning approach. *BMC Bioinformatics*, 9(1), pp.9–217.

Hu, Y.-J. et al., 2012. Decision tree-based learning to predict patient controlled analgesia consumption and readjustment. *BMC Medical Informatics and Decision Making*, 12(1), pp.12–131.

Hudson, D.L. & Cohen, M.E., 2000. *Neural networks and artificial intelligence for biomedical engineering*, New York: IEEE.

Inza, I. et al., 2010. Machine Learning: An Indispensable Tool in Bioinformatics. In R. Matthiesen, ed. *Bioinformatics Methods in Clinical Research*. Humana Press, pp. 25–48.

Johnel, O. et al., 1992. The apparent incidence of hip fracture in Europe: A study of national register sources. *Osteoporosis International*, 2(6), pp.298–302.

Keyak, J.H. et al., 1997. Prediction of femoral fracture load using automated finite element modeling. *Journal of Biomechanics*, 31, pp.125–133.

Khovanova, N., Shaikhina, T. & Mallick, K., 2014. Neural networks for analysis of trabecular bone in osteoarthritis. *Bioinspired, Biomimetic and Nanobiomaterials*, 4(1), pp.90–100.

Krikov, S. et al., 2007. Predicting kidney transplant survival using tree-based modeling. *ASAIO journal (American Society for Artificial Internal Organs : 1992)*, 53(5), pp.592–600.

Lanouette, R., Thibault, J. & Valade, J.L., 1999. Process modeling with neural networks using small experimental datasets. *Computers & Chemical Engineering*, 23(9), pp.1167–1176.

Leung, R. et al., 2013. Using a multi-staged strategy based on machine learning and mathematical modeling to predict genotype-phenotype risk patterns in diabetic kidney disease: a prospective case-control cohort analysis. *BMC Nephrology*, 14(1), pp.14–162.

Lowe, D. et al., 2013. Significant IgG subclass heterogeneity in HLA-specific antibodies: Implications for pathogenicity, prognosis, and the rejection response. *Human Immunology*, 74, pp.666–672.

More, J.J., 1978. The Levenberg-Marquardt algorithm: Implementation and theory. *Lecture Notes in Mathematics*, 630, pp.105–116.

Nguyen, D. & Widrow, B., 1990. Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. *IEEE International Joint Conference on Neural Networks*, 3, pp.21–26.

Perilli, E. et al., 2007. Structural parameters and mechanical strength of cancellous bone in the femoral head in osteoarthritis do not depend on age. *Bone*, 41, pp.760–768.

Podgorelec, V. et al., 2002. Decision trees: An overview and their use in medicine. *Journal of Medical Systems*, 26, pp.445–463.

Shaikhina, T., Khovanova, N. & Mallick, K., 2014. Artificial Neural Networks in Hard Tissue Engineering: Another Look at Age-Dependence of Trabecular Bone Properties in Osteoarthritis. In *IEEE EMBS International Conference on Biomedical & Health Informatics*. Valencia: IEEE, pp.484-487.

Sinusas, K., 2012. Osteoarthritis : diagnosis and treatment. *American Family Physician*, 85(1), pp.49–56.

Theiler, J. et al., 1992. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58, pp.77–94.

Tseng, W.-J. et al., 2013. Hip fracture risk assessment: artificial neural network outperforms conditional logistic regression in an age- and sex-matched case control study. *BMC Musculoskeletal Disorders*, 14(1), pp.14–207.

Yonaba, H., Anctil, F. & Fortin, V., 2010. Comparing Sigmoid Transfer Functions for Neural Network. *Journal of Hydrologic Engineering*, (April), pp.275–283.

Zhu, L. et al., 2013. Comparison between artificial neural network and Cox regression model in predicting the survival rate of gastric cancer patients. *Biomedical reports*, 1(5), pp.757–760.