

PARAMETRIC CONTACT MODELLING OF ROLLER AND RACE OF A ROLLER BEARING WITH NEURAL NETWORK-BASED STRESS PREDICTION

Vitalii Ovcharenko ^{1[0009-0008-1464-3425]}, Oleksii Vodka ^{1[0000-0002-4462-9869]}

¹National Technical University "Kharkiv Polytechnic Institute", 2, Kyrpychova str., 61002, Kharkiv, Ukraine
Department of mathematical modeling and intelligent computing in engineering, Part of the Educational and
Research Institute of Computer Modeling, Applied Physics and Mathematics

E-mail(s): vitalii.ovcharenko@khpi.edu.ua (✉) (corresponding author's e-mail), oleksii.vodka@khpi.edu.ua

Abstract - This study presents a parametric model of contact interaction between a roller and the inner ring of a cylindrical roller bearing, developed using the finite element method (FEM). A series of numerical experiments was conducted across a wide range of geometric and loading configurations to obtain von Mises stress distributions along the roller generatrix. The results were approximated as tabulated functions with fixed spatial discretization, enabling the formation of uniform stress vectors independent of the underlying FEM mesh. Based on the generated dataset, a regression-based neural network model was developed to directly predict stress distributions from input geometric and loading parameters. The trained model achieved high predictive accuracy on an independent test set, with a coefficient of determination $R^2 > 0.9$. An important distinction of the proposed approach from the increasingly popular physics-informed neural networks (PINNs) is the independence of the network architecture from the finite element mesh. This allows for flexible modification of the number of finite elements and geometric characteristics of the FEM model without requiring adaptation or retraining of the network, providing greater flexibility and generality for tasks such as preliminary engineering analysis or geometric optimization.

Keywords: Machine learning in engineering, Neural networks, roller bearings, Finite element analysis, Parametric modelling.

1. Introduction

The use of machine learning, particularly neural networks, in combination with the finite element method is becoming increasingly widespread in engineering applications. This enables the results of numerical studies to be used not only for evaluating the stress-strain state of systems or mechanisms with specific parameters, but also for building generalized models capable of predicting the behaviour of similar systems. In contact mechanics problems, where computations are resource-intensive, such an approach allows for rapid preliminary analysis of typical configurations without significant loss of accuracy.

One promising direction involves building direct surrogate models based on data obtained from numerical simulations, particularly finite element method (FEM) analyses. In this case, the neural network is trained on the outcomes of a series of

FEM computations and enables instant prediction of system state characteristics without the need for repeated numerical solutions.

A review of recent studies reveals various approaches for transforming finite element data into formats suitable for neural network processing. For instance, Spruegel et al. [1] proposed a comprehensive methodology for representing heterogeneous finite element analysis (FEA) data as fixed-size matrices. Their approach involves projecting the mesh and simulation results onto a spherical detector surface, followed by encoding into a structured numerical format, which enables logical validation of new simulations based on previously identified patterns.

Another research direction involves the use of neural networks for solving inverse problems. For example, Cen et al. [2] proposed a hybrid approach that combines variational methods with neural networks.

This method enables the recovery of spatially varying coefficients in diffusion equations while preserving the smoothness of the target functions and providing rigorous error estimates. Although the approach is primarily aimed at identifying hidden material properties rather than predicting stresses, it illustrates the extent to which machine learning can be integrated into traditional numerical methods.

Another example of a hybrid approach is presented by Mitusch et al. [3], where neural networks are used to augment systems of partial differential equations while strictly enforcing physical laws. Through algorithmic differentiation and implementation within the FEniCS framework, the authors were able to train neural networks that model unknown components of the governing equations – such as flow functions or spatially varying coefficients. Unlike purely statistical approaches, this methodology enables the neural network to be trained directly during the problem-solving process, ensuring a direct link between the underlying physical processes and the loss function minimization.

Despite significant progress in hybrid approaches, particularly physics-informed neural networks (PINNs), the task of direct regression remains highly relevant – that is, constructing a model that predicts stress distributions directly from geometric and loading parameters without relying on partial differential equations or intermediate physical quantities. A major limitation of PINN-based methods is their tight coupling to the structure of the finite element mesh: any change in the number of elements or domain configuration often requires redefinition of the model or retraining. In contrast, this study implements an alternative approach based on a precomputed set of FEM simulations, whose results are approximated as tabulated functions with fixed-length discretization. This enables the construction of a unified training dataset that is independent of the finite element size and mesh topology.

In the context of combining neural networks with numerical methods (including FEM), particular attention should be given to contact problems in bearing assemblies, which remain one of the most important areas in the mechanics of deformable solids. The reliability and service life of rolling-element bearings are directly influenced by the nature of interaction between rolling elements and raceways; thus, investigating such contact interactions is essential for design optimization [4].

A review of recent research in bearing technology highlights the main directions for improving the durability and reliability of rolling-element bearings. A central issue among them is the optimization of contact interactions between the rolling elements and the raceways, which has a direct impact on the overall service life of the assembly.

The main research efforts are focused on addressing several interrelated challenges. First, there is the accurate determination of loading parameters under real operating conditions, which must account for both nominal working loads and dynamic components. Modern calculation methods show that traditional simplified models often neglect substantial parasitic loads, which can reach 30–50% of nominal values.

Second, significant attention is given to analysing the influence of design features on load distribution within the bearing assembly. Comparative studies of various housing types and mounting methods have shown that rigid designs provide a more uniform load distribution among the rolling elements but may increase sensitivity to mounting misalignments. In contrast, more flexible systems offer better compliance and adaptability, though they require additional measures to protect against contamination.

One of the most promising directions is the optimization of the microgeometry of contact surfaces. Experimental studies have shown that traditional cylindrical shapes of rolling elements lead to stress concentrations at the edges of the contact zone, which is a primary cause of fatigue damage. Recent research suggests the use of specially designed profiles with smoothly varying curvature, which effectively eliminates edge effects without significantly reducing the contact area [4–7].

A major advancement in recent years has been the implementation of computational modelling techniques for solving contact problems. Comparative analysis between traditional analytical methods and modern finite element simulations has demonstrated that the latter more accurately account for real contact conditions, including load distribution irregularities, the influence of adjacent structural components, and material properties. The accuracy of such simulations reaches 90–95% when compared to experimental data.

Modern research places particular emphasis on an integrated approach to the design of bearing assemblies, combining the optimization of rolling element geometry, material properties, and lubrication systems. Studies have shown that the greatest performance gains are achieved through the simultaneous refinement of all these components, resulting in a 1.5–2 times increase in bearing service life compared to conventional solutions [9–12].

Future research is expected to focus on the development of adaptive condition monitoring systems for bearing assemblies, which would enable real-time adjustment of operating conditions and help prevent critical failures. This is especially relevant for high-load applications, where traditional maintenance strategies often prove insufficiently effective [13–18].

Alongside these developments, increasing attention has been devoted to the application of deep

learning techniques in computational mechanics. Rather than replacing classical finite element analysis, neural networks are increasingly employed to construct surrogate models that approximate the relationship between model parameters and mechanical responses, enabling rapid evaluation in design optimization, uncertainty quantification, and real-time engineering applications. Recent review papers summarize a broad spectrum of approaches, including fully connected neural networks, convolutional architectures, graph neural networks, physics-informed neural networks (PINNs), and neural operators such as DeepONet and Fourier Neural Operators (FNO), demonstrating the rapid evolution of data-driven computational mechanics [19, 23, 26].

Considerable attention has been devoted to surrogate prediction of finite element solutions, including displacement, strain, and stress fields. Representative examples include FEA-Net, which incorporates physical principles into neural network design for efficient prediction of mechanical responses [20], physics-informed frameworks capable of solving elasticity problems directly from governing equations without labelled datasets [21, 22], and neural operator architectures that learn mappings between loading conditions and complete solution fields [23–26]. More recently, DeepONet-based methods have demonstrated accurate prediction of nonlinear stress distributions for complex geometries while achieving computational speed-ups of several orders of magnitude compared with conventional finite element analysis [24, 25].

Despite these advances, most existing studies focus on predicting full-field responses on complex finite element meshes or employ sophisticated neural operator architectures designed for high-dimensional outputs. Motivated by these observations, this study proposes a compact surrogate model for predicting the contact stress profile of logarithmic roller bearings. The methodology combines a mesh-independent representation of the stress distribution with a conventional feedforward neural network, enabling rapid stress prediction while preserving the physical interpretation of the output variables.

2. Problem Setup

Rolling bearings are among the most widely used machine elements in modern engineering systems. Their reliability and operational lifetime largely depend on the nature of the contact interaction between rolling elements and raceways. During operation, external loads are transferred through relatively small contact regions, resulting in highly localized stress fields. These stresses govern the initiation and propagation of fatigue damage, surface wear, plastic deformation, and other failure mechanisms. Consequently, accurate evaluation of

contact stresses is one of the key prerequisites for the design of durable and reliable bearing assemblies.

The primary function of a rolling bearing is to transmit external loads between rotating and stationary components while minimizing frictional losses. In cylindrical roller bearings, the load is transferred through a series of contacts formed between the rollers and the raceways of the inner and outer rings. The magnitude of the load carried by each roller depends on its position within the bearing, the internal clearance, elastic deformations of the surrounding components, and manufacturing tolerances. As a result, the load is generally distributed non-uniformly among the rolling elements, leading to significant variations in local contact conditions and stress levels throughout the bearing assembly.

The theoretical foundation of most bearing contact analyses is provided by the classical Hertzian contact theory. For roller bearings, Hertzian theory provides a valuable first approximation of the contact behaviour and remains widely used in engineering calculations and bearing standards. In particular, it allows analytical determination of key contact parameters, including the half-width of the contact zone and the maximum contact pressure σ_H for cylindrical bodies under load, as shown in Eq. (1):

$$\sigma_H = \sqrt{\frac{FE^*}{\pi LR^*}} \quad (1)$$

where F denotes the applied load, L is the contact length, E^* is the effective elastic modulus, and R^* is the effective radius of curvature.

Despite their practical value, Hertzian solutions are based on a number of simplifying assumptions that are not fully satisfied in real bearing assemblies. The theory assumes perfectly smooth contacting surfaces, homogeneous isotropic materials, and ideal geometric forms. Furthermore, it neglects the influence of manufacturing tolerances, mounting errors, thermal deformations, and structural flexibility of neighbouring components. In practical applications, these factors may significantly alter the stress distribution within the contact zone and produce effects that cannot be adequately captured by analytical models. Consequently, the accuracy of Hertzian solutions gradually decreases as the geometry and loading conditions deviate from the idealized assumptions on which the theory is based.

The combined influence of geometric imperfections, elastic deformations, and assembly conditions leads not only to local stress concentrations within individual contacts but also to a complex redistribution of load throughout the bearing assembly. One of the most widely known

manifestations of this behaviour is the edge effect, where contact stresses become concentrated near the ends of the roller. At the same time, the load transmitted by different rolling elements may vary considerably depending on their position relative to the applied external force. As a result, both the local stress distribution within a single contact and the global load distribution among the rollers must be considered when evaluating bearing performance and durability.

A direct consequence of this complex contact behaviour is the non-uniform distribution of load among the rollers. Modern experimental studies and numerical simulations reveal that the actual distribution significantly deviates from classical theoretical assumptions. In particular, a pronounced non-uniformity is observed, which contradicts the idealized models of uniform loading over the upper hemisphere of the raceway and complete unloading of the lower half.

In a typical cylindrical roller bearing subjected to a radial load, the highest contact force is carried by the central roller in the upper part of the raceway, which absorbs a dominant share of the total load. In contrast, the neighbouring rollers experience substantially lower forces, which drop significantly and progressively for each subsequent element farther from the centre. This type of load distribution reflects a complex spatial interaction influenced by both design and operational features of the assembly.

Several structural and operational factors contribute to the observed load distribution within a bearing assembly. Their combined influence determines how the external load is shared between individual rolling elements and, consequently, affects the stress state of the bearing components. Understanding the relative importance of these factors is essential for the development of reliable numerical models and effective design optimization strategies.

The key factors contributing to the non-uniform load distribution include:

- Stiffness of the outer ring and housing components
- Manufacturing accuracy of the bearing elements
- Presence of mounting misalignments
- Thermal deformations during operation
- Design of the load transmission system (e.g., axle boxes, adapters, etc.)

Modern computational modelling techniques allow for detailed investigation of these effects using finite element models of the whole bearing assembly, as illustrated in Figure 1.

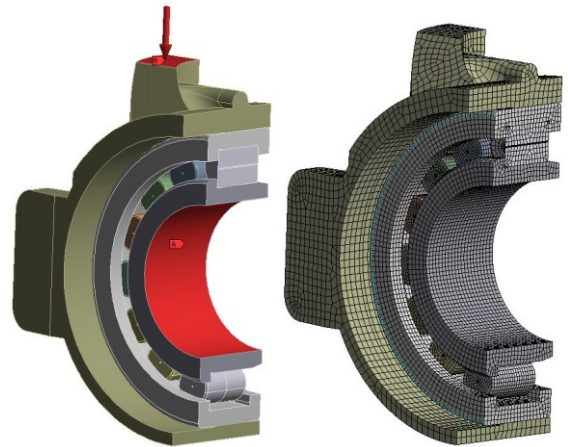


Figure 1: The model of the bearing assembly used for analysing load distribution between rollers in cylindrical roller bearings

Finite element analysis of the stress-strain state shows that deformation of the outer ring tends to form a flattened elliptical shape, leading to increased compression of the lower rollers. This effect is particularly pronounced in rigid housing designs, where deformations may reach 25–40 μm .

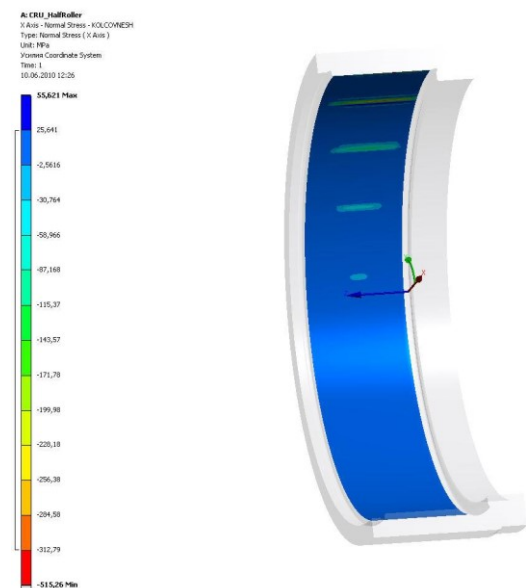


Figure 2: An example of load distribution across rollers in a cylindrical roller bearing under radial load

Load distribution optimization can be achieved through several approaches:

- Using specially profiled raceways
- Implementing adaptive load transmission systems
- Precise control of assembly clearances
- Introducing additional degrees of freedom in the housing design)

Of particular interest are comparative studies of different bearing types. For example, tapered roller bearings exhibit a more uniform load distribution (from 22% on the central roller to 1–2% on the outermost ones), which is attributed to their higher compliance and self-aligning capability. However, this comes at the cost of an increased number of contact zones, which complicates the maintenance of optimal lubrication conditions.

Modern approaches to bearing assembly design increasingly employ iterative optimization methods, where the load distribution among rollers – such as the distribution model shown in Figure 2 – serves as a key efficiency criterion. This enables a reduction of peak contact stresses by 30–40% and a corresponding increase in the service life of the assembly.

The analysis of such effects requires detailed information about the stress distribution within individual contact regions. Although analytical approaches provide valuable engineering estimates, they are generally unable to capture the combined influence of complex geometry, elastic deformations, contact nonlinearities, and varying operating conditions. As a result, finite element analysis has become one of the primary tools for investigating contact behaviour in modern bearing systems.

In practical design and optimization tasks, however, a large number of geometric and loading configurations must often be evaluated. Performing a complete finite element simulation for every candidate design may become computationally expensive, particularly when iterative optimization procedures are involved. This creates a strong motivation for developing surrogate models capable of reproducing finite element predictions with significantly lower computational cost. The present study addresses this challenge by constructing a neural-network-based model for predicting stress distributions in roller–race contacts using data generated from parameterized finite element simulations.

3. Parametric Finite Element Model and Dataset Generation

The proposed approach is based on the integration of automated finite element modelling and machine learning techniques for rapid prediction of contact stress distributions in rolling bearing assemblies. Instead of performing computationally expensive finite element analyses for each new design configuration, a comprehensive dataset is first generated using parameterized numerical simulations and subsequently employed for training a surrogate neural network model. This strategy significantly reduces computational costs while preserving the ability to capture complex stress patterns.

The overall workflow consists of several sequential stages. First, a set of geometric and loading parameters defining the contact problem is specified. Based on these parameters, a three-dimensional finite element model of the roller–race contact pair is automatically generated, meshed, and analysed. The resulting stress distribution is then extracted from the numerical solution, transformed into a standardized representation independent of the finite element mesh, and stored as part of the training dataset. This unified workflow ensures full automation and reproducibility of the data generation process.

The geometry of the roller is described using a parametric representation of its generatrix (Eq. 2), allowing systematic variation of the contact surface profile. The model includes such parameters as the roller radius, roller length, chamfer height, and the exponent controlling the shape of the profile function. By modifying these parameters, a broad range of practically relevant roller geometries can be generated while maintaining a consistent mathematical description. Such parameterization enables efficient investigation of the influence of microgeometry on the resulting contact stress distribution.

$$y = Ae^{\alpha x^N - 1} \quad (2)$$

where y is the radial distance from the cylindrical surface (defined by the fixed roller radius), x is the axial coordinate with the origin at the centre of the roller, A and α are parameters computed from the roller radius and chamfer height, and N is a variable exponent.

The parametric formulation naturally defines a multidimensional design space in which each point corresponds to a unique combination of geometric and loading conditions. The parameter space was sampled using a structured full-factorial design. Five geometric and loading parameters were varied: the roller radius R , chamfer height H , applied pressure P , roller length L , and profile exponent N . The scaling factors for the roller radius, chamfer height, and applied pressure were selected as {0.5, 0.75, 1.0, 2.0, 4.0} relative to their reference values, whereas the roller length was varied using scaling factors {0.5, 1.0}. The profile exponent N was sampled uniformly in the range 0.5–20.0 with a step of 0.5. Consequently, the complete dataset consisted of $5 \times 5 \times 5 \times 2 \times 40 = 10,000$ independent finite element simulations, each corresponding to a unique combination of input parameters. This exhaustive sampling strategy ensured a uniform coverage of the investigated parameter space and eliminated the randomness associated with stochastic sampling methods.

The finite element model is generated automatically from the specified input parameters

without manual intervention. For each parameter combination, the software constructs the three-dimensional geometry of the roller and the corresponding inner race, preserving the geometric relationships required for accurate contact analysis. Such automation enables rapid generation of numerous model variants and ensures consistency throughout the numerical study.

After the geometry has been created, the finite element model is completed by assigning material properties, generating the computational mesh, defining contact pairs, and applying the required boundary conditions and external loads. Standard contact elements are employed to model the interaction between the roller and the raceway, allowing accurate evaluation of the stress field within the contact region. The mesh density is selected automatically to provide a balance between computational efficiency and solution accuracy (see Figure 3).

All finite element analyses were performed in ANSYS Mechanical APDL using second-order 20-node SOLID186 elements for both the roller and the raceway. Surface-to-surface contact was modelled using the CONTA174 and TARGE170 contact element pair with an augmented Lagrange contact formulation. Both bodies were assumed to be made of linear elastic isotropic steel with Young's modulus $E = 210$ GPa, Poisson's ratio $\nu = 0.30$, and density 7850 kg/m^3 . The inner ring was fully constrained at its inner cylindrical surface, while symmetry conditions were imposed on the appropriate longitudinal planes of both bodies to reduce the computational domain. Automatic mesh generation was employed with local refinement in the contact region to accurately capture steep stress gradients while maintaining reasonable computational cost.

A mesh convergence study was performed before the parametric study. Meshes containing approximately 5,000, 10,000, 20,000, and 40,000 SOLID186 elements were analysed under identical loading conditions. Convergence was evaluated using both the maximum equivalent von Mises stress and the complete stress profile extracted along the roller generatrix. Increasing the mesh density from approximately 20,000 to 40,000 elements changed the maximum stress by less than 1% and the normalized stress profile by less than 0.5%, whereas the computational time increased by almost a factor of 3. Consequently, the mesh containing approximately 20,000 elements was selected for the generation of the complete dataset as the best compromise between solution accuracy and computational efficiency.

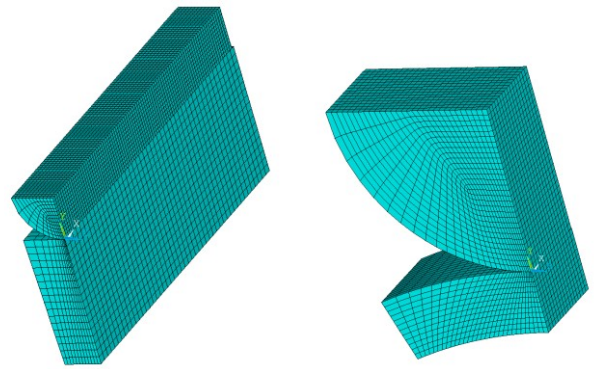


Figure 3: Finite element meshes of the roller and inner ring for contact analysis under varying input parameters

To investigate the influence of geometric parameters on contact behaviour, the entire modelling procedure is executed automatically for a large number of parameter combinations. Each simulation represents an independent numerical experiment corresponding to a unique roller geometry and loading condition. The automated workflow substantially reduces the effort required for model preparation and enables the generation of extensive datasets suitable for subsequent machine learning applications.

Following the completion of each finite element analysis, the equivalent stress distribution is extracted along the contact line between the roller and the raceway (see Figure 4). These stress values constitute the primary output of the numerical model and characterize the local contact behaviour for the corresponding parameter set. Collecting stress profiles from a large number of simulations makes it possible to analyse the effect of geometric variations and provides the foundation for constructing a predictive surrogate model.

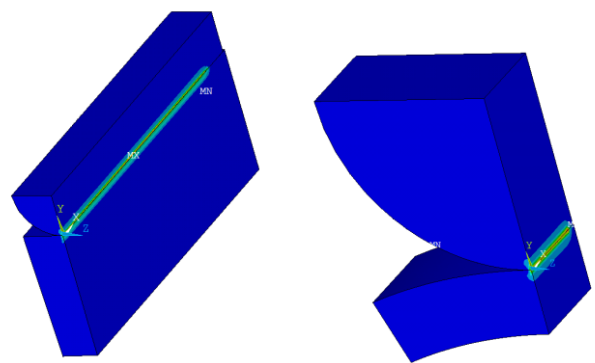


Figure 4: Equivalent stress distribution in the roller and inner ring under varying input parameters

Although the stress distribution is obtained directly from the finite element solution, its numerical representation depends on the generated mesh. Different parameter combinations may produce different mesh topologies, resulting in varying

numbers and locations of nodes along the contact line. Consequently, the extracted stress profiles cannot be directly compared or used as input for machine learning algorithms, since they do not share a common spatial discretization.

To overcome this limitation, each stress profile is transformed onto a predefined uniform spatial grid using Akima spline interpolation. This approach provides a smooth approximation of the original finite element results while preserving the characteristic features of the stress distribution, including local maxima associated with edge effects. The interpolated profile is then sampled at a fixed number of equally spaced points, producing a standardized representation independent of the underlying finite element mesh.

After interpolation, every numerical simulation is represented by a pair consisting of an input parameter vector and an output stress vector of fixed dimension. The input vector contains the geometric and loading parameters defining the analysed configuration, whereas the output vector stores the interpolated equivalent stress values at one hundred predefined locations along the contact line. Such a representation converts the finite element results into a structured dataset suitable for supervised regression.

An important advantage of the proposed data representation is its independence from mesh topology and element density. Since all stress profiles are mapped onto the same spatial coordinate system, modifications of the finite element mesh do not affect the structure of the generated dataset. This property significantly improves the flexibility of the proposed approach and allows finite element models to be refined or modified without changing the format of the training data.

The resulting dataset establishes a direct relationship between the geometric and loading parameters and the corresponding stress distribution, providing the basis for constructing a surrogate predictive model. Instead of repeatedly solving computationally expensive contact problems, the trained neural network learns this mapping from the generated examples and can subsequently predict complete stress distributions almost instantaneously for previously unseen parameter combinations.

4. Neural Network Training

The finite element model developed in this study is complemented by a neural network acting as a surrogate model capable of rapidly predicting stress distributions without performing additional numerical simulations.

The surrogate modelling problem can be formulated as the construction of a mapping between the vector of geometric and loading

parameters and the corresponding stress distribution along the roller generatrix:

$$f : x \rightarrow y \quad (3)$$

where x denotes the input parameter vector and y represents the discretized equivalent stress distribution obtained from finite element analysis.

The objective of the neural network is to approximate the computationally expensive finite element operator f_{FEM} by a significantly faster function f_{NN} , such that:

$$f_{NN}(x) \approx f_{FEM}(x) \quad (4)$$

Unlike physics-informed approaches that incorporate governing equations into the optimization process, the proposed methodology relies exclusively on a database generated by automated finite element simulations. Consequently, the neural network learns the relationship between geometry and stress response directly from numerical data while remaining completely independent of the underlying finite element mesh topology. This property allows modifications of mesh density or element type without affecting the structure or applicability of the trained model.

A critical step in constructing the surrogate model is the transformation of finite element results into a uniform numerical representation suitable for machine learning algorithms. Each simulation is characterized by a set of geometric and loading parameters that uniquely define the analysed contact configuration, while the corresponding output consists of the equivalent stress distribution along the roller generatrix.

The input vector of the neural network is defined as:

$$x = \begin{pmatrix} R \\ L \\ N \\ H \\ P \end{pmatrix} \quad (5)$$

where R is the roller radius, L is the roller length, N is the exponent defining the roller profile, H is the chamfer height, and P is the applied external load. These parameters completely determine the geometry and loading conditions of the finite element model and therefore uniquely specify the resulting contact problem.

The raw finite element results depend on the generated mesh and contain stress values at irregularly distributed nodal locations. To eliminate this dependency, the computed stress distributions are interpolated using Akima splines and subsequently resampled at one hundred uniformly

spaced positions along the roller generatrix. As a result, each numerical experiment is represented by a fixed-length output vector

$$y = (\sigma_1, \sigma_2 \dots \sigma_{100}) \quad (6)$$

where σ_i denotes the equivalent von Mises stress at the corresponding normalized coordinate. Such representation ensures that all samples possess identical dimensionality regardless of the finite element discretization, making the dataset directly compatible with standard feedforward neural network architectures.

To approximate the nonlinear relationship between the geometric parameters and the corresponding stress distributions, a fully connected feedforward neural network was adopted. The selected architecture was identified through a preliminary hyperparameter study and provides sufficient expressive capability to capture the complex dependence of contact stresses on roller geometry while maintaining low computational complexity and fast inference time.

Prior to selecting the final architecture, a systematic hyperparameter study was performed to evaluate the influence of the network topology and training parameters on predictive performance and computational efficiency. In total, about 50 network configurations were investigated. The study included fully connected architectures with one, two, and three hidden layers containing 128, 256, 512, 1024, and 2048 neurons per layer. Three activation functions (ReLU, Leaky ReLU, and GELU) were evaluated together with different learning rates (10^{-3} , 5×10^{-4} , and 10^{-4}) and batch sizes (32, 64, and 128). The models were compared using the RMSE and coefficient of determination (R^2), while also considering convergence stability, training time, inference speed, and the total number of trainable parameters. Although several deeper architectures achieved slightly lower training errors, their improvement on the independent test set was marginal and did not justify the substantial increase in computational cost or model complexity. Smaller networks, particularly those with fewer than 512 neurons, consistently exhibited reduced approximation accuracy in regions with steep stress gradients caused by edge effects. As a result, the architecture consisting of a single hidden layer with 1024 ReLU neurons was selected as the most balanced solution, providing an excellent compromise between prediction accuracy, generalization capability, computational efficiency, and implementation simplicity.

Although dimensionality reduction techniques such as principal component analysis (PCA) and autoencoder-based latent representations are frequently employed for surrogate modelling of

high-dimensional finite element fields, they were not adopted in the present study. The output of the proposed model consists of only one hundred uniformly sampled stress values describing a single one-dimensional stress profile rather than a full-field solution containing thousands of degrees of freedom. Preliminary experiments indicated that dimensionality reduction provided only negligible improvements in computational efficiency while introducing an additional approximation stage and reducing the direct physical interpretability of the predicted stresses. Therefore, the neural network was trained to predict the complete stress profile directly, preserving a one-to-one correspondence between the network outputs and the physical locations along the roller generatrix.

The adopted network receives five normalized geometric and loading parameters as input and produces one hundred output values corresponding to the interpolated stress distribution. Between the input and output layers, a hidden layer of 1024 ReLU neurons is followed by a Batch Normalization layer, which improves numerical stability and accelerates convergence by reducing the internal covariate shift. The output layer employs linear activation to predict the stress values at one hundred uniformly distributed locations along the roller generatrix (see Figure 5).

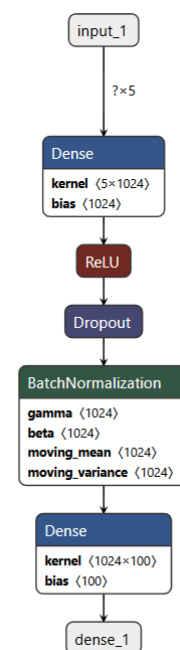


Figure 5: Architecture of the neural network

The generated dataset containing 10,000 samples was randomly divided into training and testing subsets using an 80:20 ratio. Consequently, 8,000 samples were used for network training and hyperparameter selection, whereas the remaining 2,000 previously unseen samples were reserved exclusively for evaluating the generalization capability of the final model.

Before training, all input parameters are normalized using Min-Max scaling, mapping each feature to the interval $[0,1]$. This normalization improves optimization efficiency and prevents parameters with larger numerical ranges from dominating the learning process. The resulting architecture combines simplicity with sufficient approximation capability, enabling the network to reproduce highly nonlinear stress distributions while remaining computationally efficient enough for engineering applications requiring rapid evaluation of numerous design configurations.

The available dataset was divided into training and testing subsets, with the latter reserved exclusively for evaluating the generalization capability of the model. Network parameters were optimized using the Adamax optimizer with a low learning rate to ensure stable convergence when approximating highly nonlinear stress fields. Gradient clipping was additionally employed to suppress unstable parameter updates and improve the robustness of the optimization process.

Model training was performed iteratively over multiple training stages, allowing intermediate evaluation of prediction quality and monitoring of convergence behaviour. Such an approach provides detailed information about the evolution of approximation accuracy and makes it possible to identify the point at which further training produces only marginal improvements while maintaining good generalization performance.

Prediction quality was evaluated using several complementary regression metrics. The primary optimization criterion was the root mean squared error (RMSE), defined as:

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (7)$$

where y_i and $\hat{y}_i$ denote the reference and predicted stress values, respectively.

To evaluate the overall quality of approximation independently of the stress magnitude, the coefficient of determination was calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \quad (8)$$

where $\bar{y}$ is the mean value of the reference stresses.

The simultaneous use of these metrics provides a comprehensive assessment of both absolute prediction accuracy and the ability of the neural network to reproduce the overall shape of the stress distributions.

The evolution of the evaluation metrics during training demonstrates a gradual improvement in the predictive capability of the neural network.

As the optimization proceeds, the RMSE continuously decreases, while the coefficient of determination steadily increases and eventually exceeds 0.9 on the independent testing dataset. Such behaviour indicates successful learning of the nonlinear relationship between the input parameters and the corresponding contact stress distributions.

The convergence process remains relatively slow despite the moderate number of trainable parameters (see Figure 6). This behaviour can be explained by the high sensitivity of contact stress distributions to small variations in roller geometry, particularly in the vicinity of edge effects where stress gradients become extremely steep. To ensure stable optimization under these conditions, a low learning rate together with gradient clipping was employed. Although these measures significantly reduce the risk of unstable updates and overfitting, they inevitably increase the number of training iterations required to achieve satisfactory prediction accuracy. The observed convergence characteristics therefore represent a reasonable compromise between optimization stability and computational efficiency.

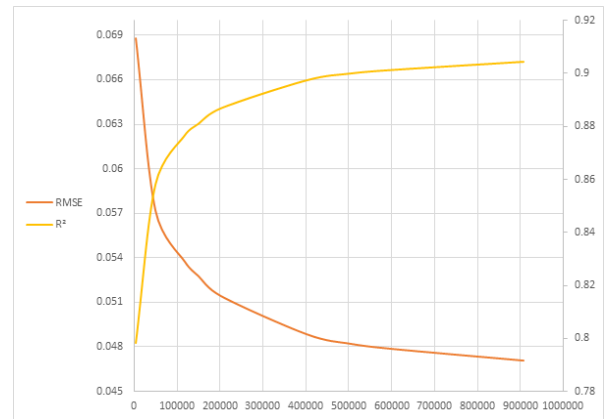


Figure 6: Evolution of RMSE and R^2 metrics during neural network training

After training, the surrogate model demonstrated the ability to reproduce the principal features of the stress distributions obtained from finite element analysis. In cases characterized by pronounced edge effects, the neural network accurately predicts both the location and magnitude of stress concentrations while preserving the overall shape of the distribution. The close agreement between the predicted and reference curves confirms that the network successfully captures the underlying nonlinear relationship between the roller geometry and the resulting contact stress field (see Figure 7).

For configurations exhibiting smoother stress distributions without pronounced local maxima, the prediction accuracy remains high, although minor deviations become more noticeable in regions where stress gradients are relatively small (see Figure 8).

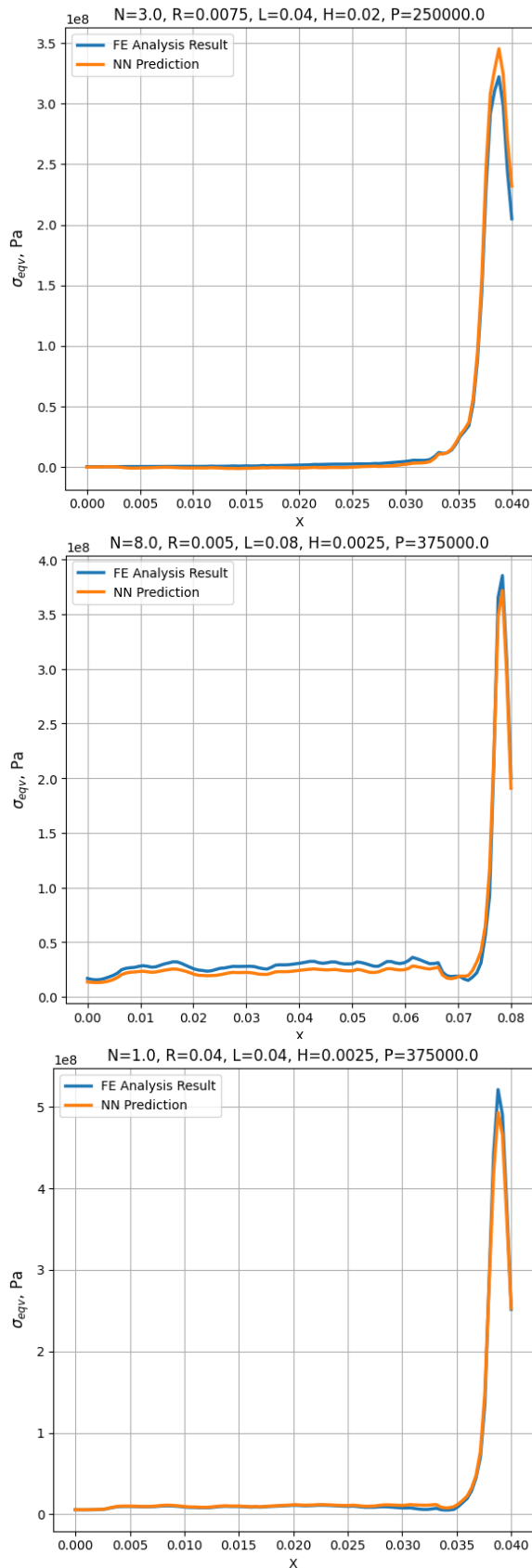


Figure 7. Predicted stress distributions in test cases with pronounced edge effects, showing the typical response captured by the trained neural network

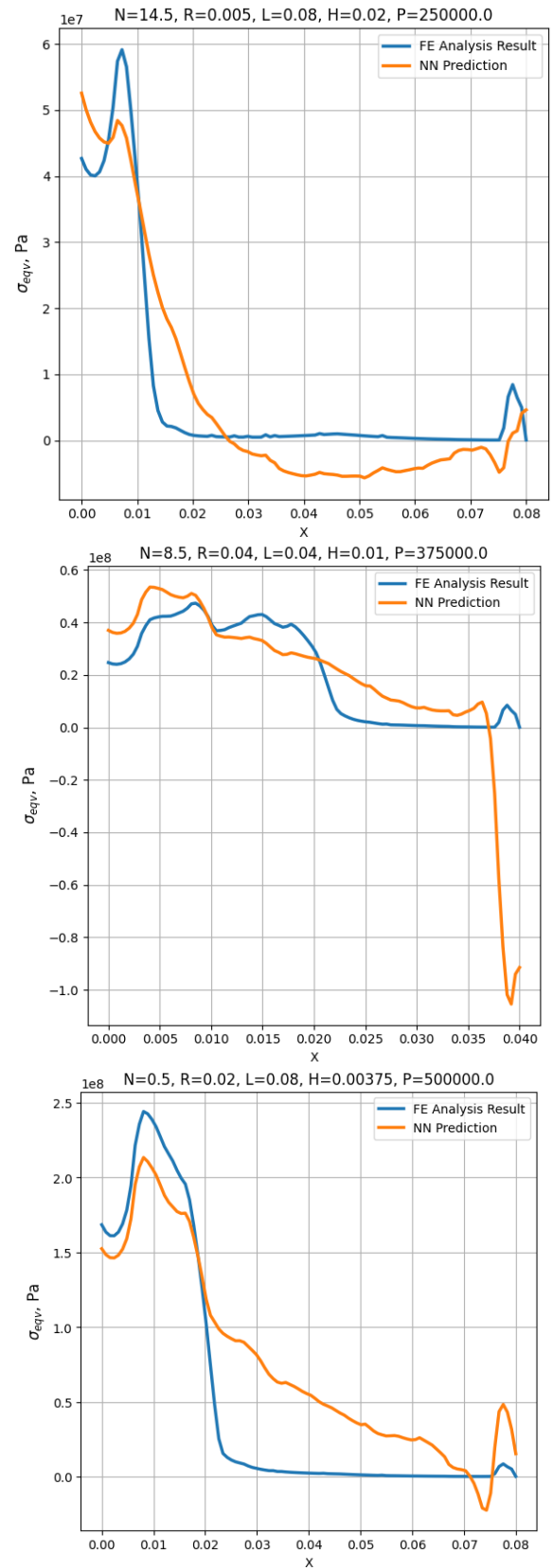


Figure 8. Stress distribution predictions for test cases without pronounced edge effects, reflecting how the neural network generalizes to smoother contact scenarios

From an engineering perspective, these discrepancies are not critical, since the overall distribution pattern and the principal stress levels are reproduced with sufficient accuracy for preliminary design analysis and optimization tasks. The results therefore demonstrate that the proposed surrogate model is capable of generalizing across a broad range of contact scenarios while maintaining consistent predictive performance.

A detailed examination of the prediction results indicates that the largest deviations occur primarily in transition regions where the stress field changes rapidly over a short spatial interval. Such behaviour is expected for regression models approximating highly nonlinear physical processes and reflects the increased complexity of representing localized stress concentrations using a finite-dimensional neural network. In contrast, regions with relatively smooth stress variation are reproduced with very high accuracy, resulting in low overall prediction errors and high values of the coefficient of determination.

From a practical standpoint, the developed surrogate model offers substantial advantages over repeated finite element simulations. Once trained, the neural network provides an almost instantaneous prediction of the complete stress distribution from a small set of geometric and loading parameters, eliminating the need for geometry generation, mesh construction, contact definition, and numerical solution. Consequently, the proposed approach can be effectively employed for preliminary engineering assessment, parametric optimization of roller profiles, sensitivity analysis, and integration into automated design workflows requiring thousands of rapid evaluations within large multidimensional parameter spaces.

5. Conclusions

A comprehensive methodology combining automated finite element modelling and machine learning techniques has been developed for predicting contact stress distributions in cylindrical roller bearing assemblies. The proposed framework enables the automatic generation of parameterized finite element models, execution of numerical simulations, and transformation of mesh-dependent stress fields into fixed-length vectors suitable for machine learning applications. This representation provides a unified dataset that is independent of the finite element discretization and facilitates the construction of efficient surrogate models.

Based on the generated database of numerical experiments, a feedforward neural network was trained to approximate the relationship between geometric and loading parameters and the corresponding equivalent stress distribution along the roller generatrix. The obtained results demonstrate that the network successfully

reproduces both strongly pronounced edge effects and smoother contact configurations, achieving a coefficient of determination exceeding 0.9 on the independent testing dataset. The close agreement between predicted and finite element stress distributions confirms the capability of the proposed model to capture complex nonlinear relationships inherent to contact mechanics problems.

An important feature of the developed approach is the complete independence of the neural network architecture from the finite element mesh topology. Unlike many hybrid or physics-informed methods, modifications of element size, mesh density, or local discretization strategy do not affect the structure of the surrogate model or require retraining. This significantly increases the flexibility of the methodology and allows the finite element model to be refined independently of the machine learning component while preserving compatibility with the existing training framework.

From a practical perspective, the developed surrogate model substantially accelerates engineering analysis by replacing computationally intensive finite element simulations with near-instantaneous neural network inference. The proposed methodology can therefore be effectively applied to preliminary design evaluation, optimization of roller microgeometry, sensitivity analysis, and automated exploration of multidimensional parameter spaces. Its computational efficiency makes it particularly attractive for engineering workflows requiring thousands of repeated evaluations.

Future research may focus on extending the parameter space to include additional geometric and material characteristics, incorporating uncertainty quantification techniques, and integrating the developed surrogate model into optimization algorithms and digital twin systems for rolling bearing assemblies. Such developments would further enhance the applicability of machine learning methods for solving complex contact mechanics problems and supporting intelligent engineering design.

References

- [1] Spruegel TC, Bickel S, Schleich B, Wartzack S. Approach and application to transfer heterogeneous simulation data from finite element analysis to neural networks. *J Comput Des Eng*. 2021;8(1):298–315. <https://doi.org/10.1093/jcde/qwaa079>
- [2] Cen S, Jin B, Quan Q, Zhou Z. Hybrid Neural-Network FEM Approximation of Diffusion Coefficient in Elliptic and Parabolic Problems. *arXiv*. 2023. [2302.10773]. <https://arxiv.org/abs/2302.10773>
- [3] Mitusch SK, Funke SW, Kuchta M. Hybrid FEM-NN models: Combining artificial neural networks

- with the finite element method. *J Comput Phys*. 2021;446:110651.
<https://doi.org/10.1016/j.jcp.2021.110651>
- [4] Harris TA. Rolling bearing analysis. New York: John Wiley & Sons; 2001.
- [5] Anatskyi YuP, Zenkevych YuA, Ovcharenko VV, Symson EA, Trokhman MV. Optymizatsiia obrazuiuchoi poverkhni rolyka podshypnyka kachennia. *Visnyk Natsionalnoho tekhnichnoho universytetu "KhPI"*. 2009;(30). Serii: Dynamika ta mitsnist mashyn.
- [6] Anatskyi YuP, Zenkevych YuA, Ovcharenko VV, Symson EA, Trokhman MV. Optymizatsiia bortiv kolets i tortsevoi poverkhni rolyka podshypnyka kachennia. *Visnyk Natsionalnoho tekhnichnoho universytetu "KhPI"*. 2009;(42). Serii: Dynamika ta mitsnist mashyn.
- [7] Symson EA, Ovcharenko VV, Shevchuk YuA. Poisk kryteriiia optymizatsii tortsevoi poverkhni rolyka i vnutrishnoi poverkhni bortiv kolets podshypnyka kachennia. *Visnyk Natsionalnoho tekhnichnoho universytetu "KhPI"*. 2010;(69). Serii: Dynamika ta mitsnist mashyn.
- [8] Ovcharenko VV. Zahalni pidkhody do proektuvannia spetsializovanoho prohramnoho kompleksu rozviazannia zadach optymizatsii. *Visnyk Natsionalnoho tekhnichnoho universytetu "KhPI"*. 2007;(38). Serii: Dynamika ta mitsnist mashyn.
- [9] Deng F, Chen Z, Liu Y, et al. A novel combination neural network based on ConvLSTM-Transformer for bearing remaining useful life prediction. *Machines*. 2022;10:1226.
<https://doi.org/10.3390/machines10121226>
- [10] Zhao W, Zhang C, Wang J, et al. Research on digital twin driven rolling bearing model-data fusion life prediction method. *IEEE Access*. 2023;11:48611–48627.
<https://doi.org/10.1109/ACCESS.2023.3277040>
- [11] Cerrada M, Sánchez RV, Li C, Pacheco F, Cabrera D, de Oliveira JV, Vásquez RE. A review on data-driven fault severity assessment in rolling bearings. *Mech Syst Signal Process*. 2018;99:169–196.
- [12] El-Thalji I, Jantunen E. A summary of fault modelling and predictive health monitoring of rolling element bearings. *Mech Syst Signal Process*. 2015;60–61:252–272.
- [13] Sobie C, Freitas C, Nicolai M. Simulation-driven machine learning: Bearing fault classification. *Mech Syst Signal Process*. 2018;99:403–419.
- [14] Lu C, Wang Z, Zhou B. Intelligent fault diagnosis of rolling bearing using hierarchical convolutional network based health state classification. *Adv Eng Inform*. 2017;32:139–151.
- [15] Saxena A, Saad A. Evolving an artificial neural network classifier for condition monitoring of rotating mechanical systems. *Appl Soft Comput*. 2007;7:441–454.
- [16] Li B, Chow M-Y, Tipsuwan Y, et al. Neural-network-based motor rolling bearing fault diagnosis. *IEEE Trans Ind Electron*. 2000;47:1060–1069.
- [17] Paya BA, Esat II, Badi MNM. Artificial neural network based fault diagnostics of rotating machinery using wavelet transforms as a preprocessor. *Mech Syst Signal Process*. 1997;11:751–765.
- [18] Moosavian A, Ahmadi H, Tabatabaeefar A, et al. Comparison of two classifiers; K-nearest neighbor and artificial neural network, for fault diagnosis on a main engine journal-bearing. *Shock Vib*. 2013;20:263–272.
- [19] Herrmann, L., Kollmannsberger, S. Deep learning in computational mechanics: a review. *Comput Mech* 74, 281–331 (2024).
- [20] Houpu Yao, Yi Gao, Yongming Liu, FEA-Net: A physics-guided data-driven model for efficient mechanical response prediction, *Computer Methods in Applied Mechanics and Engineering*, Volume 363, 2020, 112892, ISSN 0045-7825
- [21] Haghighat E., Raissi M., Moure A., Gomez H., Juanes R., A deep learning framework for solution and discovery in solid mechanics, 2020, <https://arxiv.org/abs/2003.02751>
- [22] Lu Lu, Xuhui Meng, Shengze Cai, Zhiping Mao, Somdatta Goswami, Zhongqiang Zhang, George Em Karniadakis, A comprehensive and fair comparison of two neural operators (with practical extensions) based on FAIR data, 2021, <https://arxiv.org/abs/2111.05512>
- [23] Yin M, Zhang E, Yu Y, Karniadakis GE. Interfacing finite elements with deep neural operators for fast multiscale modeling of mechanics problems. *Comput Methods Appl Mech Eng*. 2022 Dec 1;402:115027. doi: 10.1016/j.cma.2022.115027. Epub 2022 May 10. PMID: 37384215; PMCID: PMC10300559.
- [24] Jangseop Park, Namwoo Kang, Point-Deeponet: Predicting nonlinear fields on non-Parametric geometries under variable load conditions, *Neural Networks*, Volume 198, 2026, 108560, ISSN 0893-6080
- [25] Jha, P.K. From Theory to Application: A Practical Introduction to Neural Operators in Scientific Computing. *Mathematics* 2026, 14, 2421.
- [26] B.P. van de Weg, L. Greve, M. Andres, T.K. Eller, B. Rosic, Neural network-based surrogate model for a bifurcating structural fracture response, *Engineering Fracture Mechanics*, Volume 241, 2021, 107424, ISSN 0013-7944