ISSN 2579-7069

Exploring Transformer Life Forecasting through an In-Depth Analysis Utilizing the Random Forest Algorithm in Research and Development

Lei Gan[1,*], Hao Wu[2], Manal A. Ismail[3]

[1]*School of Science, Harbin Institute of Technology, China*

[2]*School of Aerospace Engineering and Applied Mechanics, Tongji University, China*

[3]*Faculty of Engineering, Helwan University, Helwan, Cairo, Egypt*

(Received: October 18, 2023; Revised: November 20, 2023; Accepted: December 18, 2023; Available online: January 7, 2024)

Abstract

Accurately assessing the life and operating status of transformers has important guiding significance for the formulation of maintenance strategies for power grid companies, and at the same time plays a key role in the risk management of power grid companies. However, the traditional methods for predicting the remaining life of the equipment have the problems of insufficient accuracy or long data training time. In order to achieve a more accurate assessment of the life and status of the transformer, a random forest-based transformer life prediction method is constructed in this paper. Relying on the theory of big data analysis, by mining and analyzing the accumulated data of massive transformers, the life prediction model of the transformer is established and the characteristic parameters affecting the life of the transformer are extracted to predict the life of the transformer. The experimental data research demonstrates that the model can be accurate and effective Predicting the life of transformers has higher prediction accuracy than traditional methods, providing method references for asset management and risk management of power grid companies.

Keywords: Predicted Lifetime, Random Forest, Transformer, Prediction Accuracy

1. Introduction

Power transformer is a key asset in power grid equipment and plays a vital role in the reliability of power system operation. During the operation and maintenance period, the investment of overhauls technical innovation project would not only affect the safe and stable operation of the power grid, but also affect the profitability output of the company. Based on the goal of technical feasibility and the best economic benefits, realizing the precise decision of capital investment in equipment overhauls technical innovation project was an important research direction at the moment [1]. At the same time, the continuous increase in energy demand and the increase in the number of operating transformers have made the operating transformers close to or have exceeded their expected technical life, leading to the problem of high failure rates of operating transformers [2]. The failure of an in-service transformer will have catastrophic consequences for the economy and operation of the power grid. Therefore, it is necessary to conduct regular condition monitoring of the transformer to predict the life of the transformer to plan equipment maintenance and replacement and reduce the risk of equipment failure. How to use the advantages of big data to find a data-supported analysis method for the life management of transformers is a hot research issue for smart grids [3][4].

The main purpose of data mining technology is to explore the relationship between hidden variables from big data sets. Data mining technology involves three aspects: statistical learning, artificial intelligence and machine learning. In addition, data mining techniques are also used to analyze and predict research objects. Most of the data mining

***Corresponding author: Lei Gan (leigan.wu@tongji.edu.cn)**
DOI: https://doi.org/10.47738/ijiis.v7i1.192

techniques used for analysis use clustering algorithms and association rules, while the data mining techniques used for prediction mainly use classification and regression algorithms. These include decision trees, artificial neural networks, genetic algorithms, K nearest neighbors, and naive Bayes. Generally speaking, the process of data mining on large data sets includes 7 steps. These steps can be defined as "data cleansing", "data integration", "data selection", "data conversion", "data mining", "model evaluation" and "analysis report".

So far, most of the research work in this field has focused on the discussion of the first two procurement strategies, operation and maintenance strategies, and LCC cost calculation methods, while the research on how to formulate decommissioning strategies, especially determining the economic decommissioning point of equipment, is very limited. Literature 5 proposes a method to determine the economic retirement point of the main equipment of the power grid by constructing an equipment failure model [5]. A large number of researchers have used data mining technology in power-related fields [6], For example, in document 6 and 7 [7].Applying the random forest algorithm to the field of power load has ideal results. Use big data processing to mine the relevant variables of each transformer, find out how it affects the life of the transformer, and predict the life of the transformer.

Some scholars have proposed to use the massive data of electricity meters for correlation analysis, and use data mining techniques to find out the hidden relationship between big data and performance status of electricity meters that are difficult to explore with traditional statistical methods. In literature 8, the characteristic parameters that affect the life of the transformer are extracted, and these characteristic parameters are learned through an adaptive fuzzy neural network, and the back propagation algorithm is used to solve the adaptive dynamic adjustment of the weights, and the life prediction model of the transformer is constructed [8]. In Literature 9, a dynamic failure rate model of a transformer suitable for short-term and medium- and long-term prediction is established based on the Markov model, and the remaining life of the transformer is modeled according to the calculated failure rate [9]. In Literature 10, a deep belief network is used to extract and classify the multi-dimensional data of power transformer faults, and combined with D-S evidence theory to solve the uncertainty problem in fault diagnosis, a multi-level decision fusion model for power transformer fault diagnosis is constructed [10]. However, the above methods have problems such as insufficient prediction accuracy and long data training time.

This paper proposes a transformer model based on random forest (RF). By mining and analyzing the basic attribute information of the transformer, the service life of the transformer is used as the output label of the model to predict [11]. Each meter can be obtained before the transformer is put into use. The estimated life of the transformer provides an analysis basis for the monitoring and rotation cycle of the transformer, and verifies the accuracy of the model proposed in this article. The transformer life prediction based on the random forest algorithm has extremely high accuracy and is currently a new prediction method in terms of transformer life prediction. It has great advantages over the previous prediction methods and has improved prediction accuracy and training duration.

2. Literature Review

Random forest is one of the extension methods of decision trees. It is an integrated learning algorithm composed of multiple decision trees [12][13]. Because decision trees are prone to overfitting, in order to improve this shortcoming, the prediction result of random forest is composed of multiple decision trees. The trees are independently voted on, and the combination of decision trees makes parallel training of data sets possible. When the data set is large in scale and complex in nature, a single decision tree is not enough to obtain the data information in the synchronized phasor data, and a single tree requires more time to classify the entire data set, and multiple decision trees are used to work in parallel. The speed and accuracy of classifying data sets are very efficient. The classification principle of random forest is shown in Figure 1.

Random forest is a powerful machine learning algorithm that deviates from conventional decision tree construction. Instead of employing all variables to split tree nodes, it adopts a distinctive approach by selecting a random subset of variables at each node for determining the optimal split. This deliberate randomization serves a crucial purpose: to mitigate the correlation between decision trees within the forest, thereby reducing the overall variance of the ensemble. The construction process of a random forest typically involves several key steps, designed to harness the benefits of this randomness. These steps encompass the careful selection of variable subsets, the creation of individual decision trees

through iterative node splitting, and the amalgamation of these diverse trees into a robust and resilient forest. By introducing controlled randomness into the decision-making process, random forest emerges as a robust ensemble learning technique, capable of enhancing predictive accuracy and generalizability across various applications.

1) Extract n-tree sample subsets from the original data. To enhance the clarity and depth of the research, the initial step involves extracting n-tree sample subsets from the original dataset. This crucial phase is undertaken to obtain representative and diverse samples that encapsulate the inherent complexity and variability present in the entire dataset. By breaking down the data into subsets, we aim to create a more manageable and focused set of observations for analysis. This process not only facilitates a more nuanced understanding of the dataset but also serves as a foundation for subsequent analytical procedures. The careful extraction of these subsets lays the groundwork for a comprehensive exploration of patterns, trends, and relationships within the data, ultimately contributing to the robustness and reliability of the research outcomes.
2) The research approach involves the utilization of sample subsets to construct decision trees. In this process, each subset is systematically employed to generate a decision tree, wherein at each node of the tree, a variable denoted as M is randomly chosen to facilitate the splitting of the data. The objective is to expand the tree iteratively, ensuring that the terminal nodes attain a minimum size threshold, thereby avoiding the creation of terminal nodes with a number of instances falling below a specified threshold. This strategy aims to enhance the robustness and generalizability of the decision trees by promoting a sufficient level of granularity in the final nodes. The randomness in variable selection adds an element of diversity to the decision tree construction, contributing to the overall stability and reliability of the generated models. The approach emphasizes the importance of balancing tree growth with node size, aligning with the objective of producing decision trees that effectively capture the underlying patterns and relationships within the data.
3) Use voting mechanism to count the results of n-tree decision trees for classification. Random forest adopts the mode of multiple decision trees working in parallel, sampling the data randomly with replacement, and its predictive ability is relatively better than the single classification model. It is suitable for large data sets. Its classification model is generally considered to have high precision.

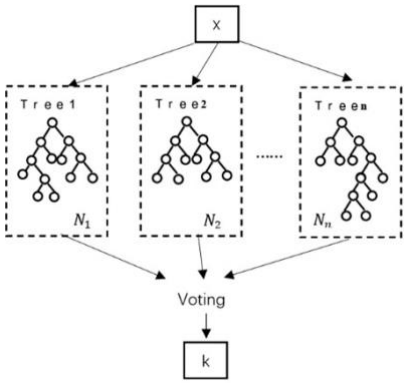

Figure 1. Random forest classification principle

3. Research Methodology

This paper preprocesses the collected transformer data and divides it into a training set and a test set, and then filters out multiple variables that may be associated with the life of the transformer, establishes a random forest model through computer programming and training set data, and substitutes it into the test set Data and adjust parameters to obtain different prediction results, and compare with the prediction results of the SVM method [14][15].

3.1. Data Preprocessing

The transformer data used in this article is provided by Anhui Power Grid. The transformer data in the data includes 14 basic information for analysis, including: "equipment classification", "start date", "asset manufacturer", "model", " Asset manufacturing country", "maintenance factory", "maintenance factory", "factory area", "cost center", "physical management department", "use storage department", "use custodian", "voltage level", "equipment Attribute data of

various types of transformers: increase method" and "transformation capacity". These data are converted into corresponding digital labels, which are respectively used as the input feature quantities of the prediction model, and the life expectancy can be predicted by establishing the model.

When constructing the equipment life prediction model, the service life of 2000 transformers is counted and classified according to the service life, which is divided into "A", "B", "C", "D", "E" There are 7 levels of "F" and "G", where A level represents the equipment service life of 0-5 years, the B level represents the equipment service life of 5-10 years, and the C level represents the equipment service life of 10-15 years. D Class E means the service life of the equipment is 15-20 years, Class E means the service life of the equipment is 20-30 years, Class F means the service life of the equipment is 30-40 years, and Class G means the service life of the equipment is more than 40 years. The remaining 14 items of basic attribute information are combined and converted into input feature quantities composed of 15 numbers.

3.2. Establishment of the Life Prediction Model of Transformer Equipment

This research involves an extensive dataset comprising over 2,065 electricity meters, serving as a fundamental aspect of the investigation. Notably, the dataset is divided into two subsets for distinct purposes. A substantial portion, precisely 2000 data points, is allocated for training the model. This training dataset plays a crucial role in facilitating the learning process and enhancing the model's predictive capabilities. Additionally, to assess the model's efficacy and generalization, 65 pieces of data are reserved for testing purposes. These data points serve as a predictive benchmark, allowing for the evaluation of the model's performance in real-world scenarios [16][17]. The specifics of the prediction process are visually represented in Figure 2, providing a comprehensive overview of the conceptual framework guiding the study. This structured approach to data utilization ensures a rigorous evaluation of the model's predictive accuracy and its potential applicability in practical electricity metering scenarios.

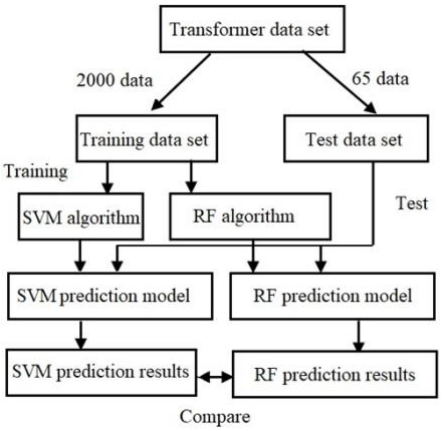

Figure 2. Flow chart of transformer life prediction

In the course of this research, the gathered equipment data undergoes a meticulous preprocessing stage aimed at extracting essential features, ultimately resulting in the creation of a comprehensive transformer dataset encompassing various influential factors [17][18]. This dataset is subsequently partitioned into two distinct segments for the purpose of model development and evaluation. The first segment, comprising 2000 sets of data, serves as the training dataset, facilitating the refinement and optimization of the models. Concurrently, the second segment, consisting of 65 sets of data, is earmarked for testing the models, thereby assessing their predictive capabilities.

Subsequently, a rigorous analysis is conducted on the predictions generated by the two distinct models developed during the study. This analytical process entails a thorough examination of the models' performance across the testing dataset. By comparing and contrasting the outcomes produced by each model, the research aims to derive insightful conclusions regarding their respective efficacy and accuracy in predicting outcomes based on the equipment data. The final stage of the investigation involves synthesizing these comparative analyses to draw overarching conclusions that contribute to a deeper understanding of the models' performance and their applicability in the context of the equipment dataset under consideration [19][20].

 ISSN 2579-7069

4. Result and Discussion

To assess the robustness and reliability of the prediction model proposed in this article, a comprehensive experimental validation is conducted on the transformer data. This validation process involves meticulous comparisons with previously retained data, enabling a thorough examination of the model's predictive accuracy and performance consistency. The verification method employed is designed to scrutinize the reliability of the model, ensuring that it can effectively generalize to new data and maintain its predictive efficacy across different scenarios. By subjecting the proposed prediction model to rigorous experimental verification, this study aims to contribute valuable insights into its practical applicability and establish a foundation for its confident adoption in real-world scenarios. The integration of empirical evidence and systematic validation procedures enhances the credibility of the proposed model, reinforcing its potential as a reliable tool for making accurate predictions in diverse contexts.

4.1. Transformer Life Prediction Results

In the life prediction model of the electric meter, the study employs a comprehensive approach by incorporating 14 distinct feature vectors as input parameters. The evaluation of prediction accuracy is visualized through Figure 3, which illustrates the error associated with the prediction outcomes. To establish the robustness and reliability of the model, a dataset comprising 2000 transformer equipment records is utilized for the training phase, while an additional set of 65 equipment records is reserved for testing purposes.

The experimentation process involves varying the number of decision trees, denoted as n-tree values, during the model training. This exploration allows for a nuanced analysis of the model's performance under different conditions. The ensuing results, meticulously presented in the accompanying figure, shed light on the impact of varying n-tree values on the predictive capabilities of the electric meter's life prediction model. This systematic investigation aims to uncover optimal configurations and parameters that enhance the overall accuracy and reliability of the prediction model, contributing valuable insights to the field of electric meter life prediction.

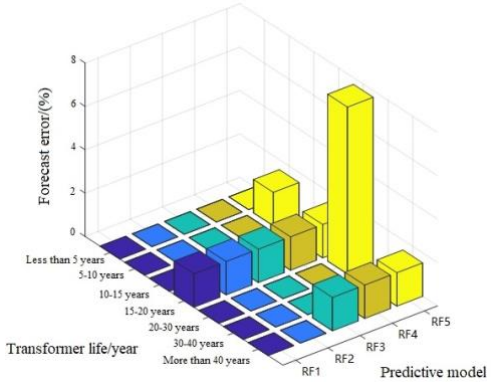

Figure 3. RF prediction error map of different transformer life classes

The provided information outlines the number of trees (n-tree values) associated with different random forests (RF1, RF2, RF3, RF4, and RF5). These values, namely 25, 20, 15, 10, and 5, respectively, signify the quantity of decision trees within each random forest. This configuration is crucial in understanding the structure and complexity of the random forests under consideration.

The variation in the number of trees among the random forests suggests potential differences in their predictive power and generalization capabilities. Typically, a higher number of trees can enhance the robustness and accuracy of a random forest model, as it combines predictions from multiple trees to mitigate overfitting and improve overall performance. Conversely, a lower number of trees may result in a simpler model that is computationally more efficient but may sacrifice some predictive accuracy.

The observed gradient in the n-tree values across the random forests prompts further investigation into the rationale behind this selection. Researchers may explore whether these specific quantities were chosen based on optimization processes, computational constraints, or if there is a deliberate design strategy underlying the varying tree numbers. Additionally, understanding the impact of these n-tree values on the model's performance could be a key focus, shedding

 ISSN 2579-7069

light on the trade-offs between model complexity and predictive accuracy within the context of the specific problem being addressed.

It can be seen from the table that using the random forest prediction model constructed in this article, the accuracy of the equipment lifetime predicted by the basic information of 15 transformer equipment is above 95% except for the experimental result with an n-tree value of 5. And the larger the n-tree value, the lower the prediction error, and the better the prediction effect, and there are a small amount of prediction error for transformers with a service life of 15-20 years. However, the training time required for prediction will continue to increase with the increase of the n-tree value. If the amount of data is large, a lower n-tree value can be used for life prediction to balance the prediction time and the prediction effect. According to different needs make changes.

4.2. Random Forest Prediction Vs. SVM Prediction

Similarly, use the SVM algorithm to predict the life of the equipment using the data of 2000 transformer equipment and 65 equipment data. The results are shown in Figure 4 below. In most life classes, the number of scrapped equipment predicted by the SVM algorithm is greater than the number of true scraps, and the predicted number of scrapped equipment for the transformer life class of some equipment is less than 50% of the real data.

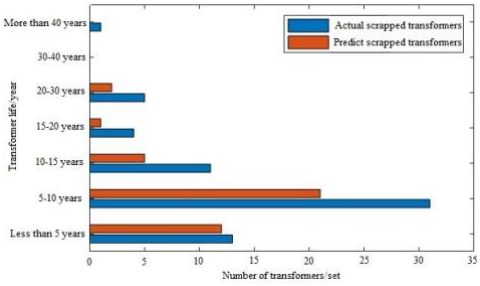

Figure 4. The result of life prediction of transformer equipment based on SVM algorithm

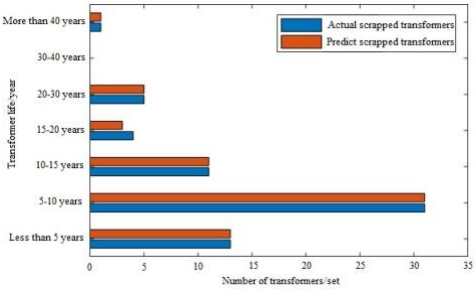

Figure 5. The life prediction results of transformer equipment under the random forestmodel

In Figure 5, the prediction result of the random forest prediction model with an n-tree value of 25, except that there are a small amount of error in the prediction of the life level of 15-20 years, the prediction accuracy of the rest of the life level is 100%, and the overall prediction is accurate The rate is 98.45%, indicating that random forest can predict the life of equipment well based on the given data. The comparison between the prediction results of the random forest prediction model with an n-tree value of 25 and the life prediction results of transformer equipment based on the SVM algorithm is shown in Figure 6 below.

It can be seen from the figure that the prediction accuracy of the random forest model under the transformer life level is higher than that of the SVM model, except that the sample number of the equipment life level of 30-40 years is 0. It shows that the prediction accuracy of the random forest model is higher than that of the SVM model. Under the condition of 15 items of data, the prediction effect of the random forest model in the prediction of transformer life is much higher than that of the SVM model. Moreover, the data training time of the SVM model is much longer than that of the random forest model. From the overall accuracy analysis, the accuracy of the SVM model is 63.08%, which is much lower than the 98.45% of the random forest model.

 ISSN 2579-7069

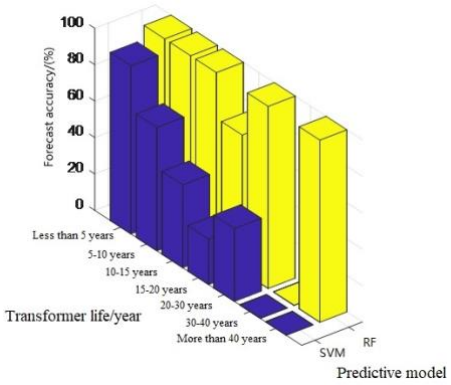

Figure 6. Comparison of prediction results between RF model and SVM model

5. Conclusion

This research leverages the capabilities of the random forest algorithm to develop a transformative life prediction model for transformers. The study involved an extensive experimental test using a dataset comprising 2065 transformer records. The outcomes of the experiment yield significant insights into the effectiveness of the proposed model:

1) The random forest model, as constructed in this paper, demonstrates a remarkable ability to predict transformer life accurately. Comparative analysis against alternative algorithms substantiates the model's superior prediction accuracy, highlighting its robust performance.

2) Investigation into the model's parameter, n-tree, reveals a noteworthy trend. As the value of n-tree increases, so does prediction accuracy; however, this comes at the expense of increased prediction time. To strike a balance between accuracy and efficiency, it is recommended to adjust the value of n-tree based on specific needs and considerations.

3) The methodology adopted in this research aligns with big data principles, emphasizing simplicity and practicality in its implementation. This approach holds promise for streamlining asset management and enhancing risk management practices within power grid enterprises. Its compatibility with the overarching concept of big data ensures a user-friendly and effective solution for industry professionals. Overall, this research contributes valuable insights and a practical tool for optimizing transformer life predictions in the context of power grid operations.

6. Declarations

6.1. Author Contributions

Conceptualization: L.G., H.W., and M.A.I.; Methodology: H.W.; Software: L.G.; Validation: L.G., H.W., and M.A.I.; Formal Analysis: L.G., H.W., and M.A.I.; Investigation: L.G.; Resources: M.A.I.; Data Curation: M.A.I.; Writing Original Draft Preparation: L.G., L.G., and M.A.I.; Writing Review and Editing: M.A.I., L.G., and H.W.; Visualization: L.G.; All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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

References

[1] Li Na, Wang Xiaoliang, Li Chengqi, Zhang Zhenjun, Zhang Weiwei, "The overhauls technical innovation project optimization method of power grid device based on Life Cycle Asset Management," *Energy Reports*, vol. 6, no. S9, pp. 1-8, 2020.

[2] L. Xue, Y. Liu, Y. Xiong, Y. Liu, X. Cui, G. Lei, "A data-driven shale gas production forecasting method based on the multi-objective random forest regression," *Journal of Petroleum Science and Engineering*, vol. 196, Art. no. 107801, pp. 1-13, 2021.

[3] Wu Moxuan, "Analysis of assets' life cycle cost and benefits of technical overhaul," *Jiangxi Electric Power*, vol. 1, no. 1, pp. 77–94, 2017.

[4] Liang Gang, Li Shengwei, Guo Tiejun, et al., "Assistant decision-making method for transformer-replacement based on equivalent annual cost in life cycle," *Proc CSU - EPSA*, vol. 29, no. 6, pp. 130–4, 2017.

[5] XIE Ning, WANG Chengmin, XIAO Dingyao, SONG Xiaonan, "Economic Disposal Time of Primary Electricity Equipments," *Electric Power Construction*, vol. 35, no. 06, pp. 165-168, 2014.

[6] Da Liu, Kun Sun, "Random forest solar power forecast based on classification optimization," *Energy*, vol. 187, no. 1, pp. 1-10, 2019.

[7] Huang Jitao, Fan Bo, Zhou Yuanfeng, Hu Tingting, Liang Fei, Zeng Xiaodong, "Fault and life prediction model of smart meter based on random forest," *Ordnance Industry Automation*, vol. 38, no. 10, pp. 57-60, 2019.

[8] Hu Biwei, Deng Xiangli, Jia Shenghao, "Transformer life estimation and state assessment based on ANFIS," *Electrical Measurement & Instrumentation*, vol. 1, no. 1, pp. 1-8, Feb. 8, 2021.

[9] F. Lo, C.M. Bitz, J.J. Hess, "Development of a Random Forest model for forecasting allergenic pollen in North America," *Science of The Total Environment*, vol. 773, Art. no. 145590, pp. 1-13, 2021.

[10] E. Mussumeci, F. Codeço Coelho, "Large-scale multivariate forecasting models for Dengue - LSTM versus random forest regression," *Spatial and Spatio-temporal Epidemiology*, vol. 35, Art. no. 100372, pp. 1-12, 2020.

[11] J. Keränen, P. Saarinen, and V. Hongisto, "Prediction accuracies of Ray-tracing and regression models in open-plan offices," *Building and Environment*, vol. 239, no. 1, pp. 110406–110413, 2023. doi:10.1016/j.buildenv.2023.110406

[12] S. Li et al., "How spatial features affect urban rail transit prediction accuracy: A deep learning based passenger flow prediction method," *Journal of Intelligent Transportation Systems*, vol. 1, no. 1, pp. 1–12, 2023. doi:10.1080/15472450.2023.2279633

[13] H. Wang and P. Budsaratragoon, "Exploration of an "Internet+" Grounded Approach for Establishing a Model for Evaluating Financial Management Risks in Enterprises", *Int. J. Appl. Inf. Manag.*, vol. 3, no. 3, pp. 109–117, Sep. 2023.

[14] T.-C. T. Chen, H.-C. Wu, and M.-C. Chiu, "A deep neural network with modified random forest incremental interpretation approach for diagnosing diabetes in Smart Healthcare," *Applied Soft Computing*, vol. 1, no. 1, pp. 111183–111190, 2023. doi:10.1016/j.asoc.2023.111183

[15] A. P. Wibawa et al., "Mean-Median Smoothing Backpropagation Neural Network to Forecast Unique Visitors Time Series of Electronic Journal," *J. Appl. Data Sci.*, vol. 4, no. 3, pp. 163–174, Aug. 2023

[16] M. Jiang, J. Wang, L. Hu, and Z. He, "Random Forest clustering for discrete sequences," *Pattern Recognition Letters*, vol. 174, no. 1, pp. 145–151, 2023. doi:10.1016/j.patrec.2023.09.001

[17] Y. Shi, "Formulation and Implementation of a Bayesian Network-Based Model", *Int. J. Appl. Inf. Manag.*, vol. 3, no. 3, pp. 101–108, Sep. 2023.

[18] M. L. Santella et al., "Predicting the creep-rupture lifetime of a cast austenitic stainless steel using Larson-Miller and Wilshire parametric approaches," *International Journal of Pressure Vessels and Piping*, vol. 205, no. 1, pp. 105006–105013, 2023. doi:10.1016/j.ijpvp.2023.105006

[19] S. N. Maharani, B. Sugeng, M. Makaryanawati, and M. M. Ali, "Bank Soundness Level Prediction: ANFIS vs Deep Learning," *J. Appl. Data Sci.*, vol. 4, no. 3, pp. 175–189, Sep. 2023,

[20] S. Han et al., "A deep neural network approach combined with Findley parameter to predict fretting fatigue crack initiation lifetime," *International Journal of Fatigue*, vol. 176, no. 1, pp. 107891–107898, 2023. doi:10.1016/j.ijfatigue.2023.107891