Hindawi
Computational Intelligence and Neuroscience
Volume 2021, Article ID 6049195, 6 pages
https://doi.org/10.1155/2021/6049195

Research Article

Enterprise Risk Assessment Based on Machine Learning

Boning Huang,[1] Junkang Wei,[2] Yuhong Tang,[3] and Chang Liu[4]

[1]*Shenzhen University Webank Institute of Fintech, Shenzhen University, Shenzhen 518052, China*
[2]*School of Pharmaceutical Sciences, Sun Yat-sen University, Guangzhou 510630, China*
[3]*School of Business and Tourism, Sichuan Agricultural University, Chengdu 610000, China*
[4]*Department of Qualitative Economics and Mathematics, School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China*

Correspondence should be addressed to Chang Liu; z0004574@zuel.edu.cn

Received 24 October 2021; Accepted 5 November 2021; Published 16 November 2021

Academic Editor: Bai Yuan Ding

Scientific risk assessment is an important guarantee for the healthy development of an enterprise. With the continuous development and maturity of machine learning technology, it has played an important role in the field of data prediction and risk assessment. This paper conducts research on the application of machine learning technology in enterprise risk assessment. According to the existing literature, this paper uses three machine learning algorithms, i.e., random forest (RF), support vector machine (SVM), and AdaBoost, to evaluate enterprise risk. In the specific implementation, the enterprise's risk assessment indexes are first established, which comprehensively describe the various risks faced by the enterprise through a number of parameters. Then, the three types of machine learning algorithms are trained based on historical data to build a risk assessment model. Finally, for a set of risk indicators obtained under current conditions, the risk index is output through the risk assessment model. In the experiment, some actual data are used to analyze and verify the method, and the results show that the proposed three types of machine learning algorithms can effectively evaluate enterprise risks.

1. Introduction

With the development of artificial intelligence and the advent of the era of big data, many scholars have used machine learning methods to conduct extensive research on risk assessment [1–4]. Enterprise risk management plays an important role in the stable operation of financial institutions at home and abroad. The traditional methods of judging whether users are in default can no longer meet the requirements of today's multiple types of data, large number of users, and high risk prediction accuracy [5–7]. A large number of scholars use machine learning methods. In-depth related discussions and a series of research results have been made to prove that the method has good prediction and generalization capabilities [8–10].

In the early days, researchers mainly used risk assessment methods based on statistical learning methods. Methods such as regression analysis were first used in the field of credit risk assessment. The linear discriminant analysis method was used for the credit evaluation system, and a mathematical statistics-based model was built to study the credit risk evaluation problem [11–14]. However, these methods have certain limitations. It is too hypothetical for the data distribution requirements, and the sample classification is based on the variance instead of the mean, so the final classification effect is not particularly strong. Linear regression was used to make a score rating based on the credit status of the lender and actual situation [15–18] to forecast the credit risk of bank customers. In essence, the linear regression method uses the existing user credit data to perform regression prediction on users with unknown credit status and finally obtains the probability of whether the user defaults. However, the linear regression also has certain drawbacks [19–22]. The value range it obtains is between plus and minus infinity, and the emergence of logistic regression has just solved this problem. Wiginton et al. first proposed the logistic regression model for credit evaluation [20]. Logistic regression uses the sigmoid function to convert

the value obtained after linear regression into a probability value and sets an empirical threshold between 0 and 1 to realize the binary classification problem [23–25].

The risk assessment model based on machine learning has gradually emerged in recent years, showing its superiority compared with traditional risk assessment methods. Common modern machine learning methods include BP neural network, K nearest neighbors (KNN), support vector machine (SVM), etc. In addition, the machine learning methods based on tree models are also widely used in risk assessment, such as basic decision tree models and integrated models such as random forest (RF), GBDT, XGBoost, and LightGBM. Makowski first used modern machine learning methods for risk assessment, in which the credit data were employed to build a model on the classification tree to classify good and bad customers. KNN was also validated more efficiently for two-class classification problem. The artificial neural network model was applied to the personal credit scoring model, which constructed a scoring system based on user credit data. The experimental results show that ensemble models such as RF perform very good in risk assessment. Some researchers pointed out that the machine learning method is mainly to model the historical risk data through supervised learning. After a series of operations such as data processing and feature extraction, the constructed model is used to predict user behavior and characteristics to determine the enterprise risk.

According to the existing literature, this paper uses machine learning algorithms for enterprise risk assessment. Specifically, three types of representative machine algorithms: RF, SVM, and AdaBoost, are used to analyze and evaluate the risk of a certain company. Based on the establishment of a corporate risk indicator system, three types of machine learning algorithms are trained using corporate historical data to obtain a reliable evaluation model. On this basis, the current state of the enterprise is evaluated and judged, and its risk assessment results are obtained. In the experiment, actual data are used to test and evaluate the performance of the three types of machine learning algorithms, and the results show their effectiveness and reliability.

2. Index for Evaluation of Enterprise Risk

The risk status of the enterprise directly determines the borrower's ability and willingness to repay the loan with cash flow. Therefore, it is necessary to establish a scientific and intuitive indicator system to provide support for bank loan decision making, so as to make evaluations scientifically and objectively. For example, in the analysis of factors affecting credit decision making, it should comprehensively consider the various influencing factors of credit risk. According to the previous research studies, this paper uses the seven evaluation indicators to describe the enterprise risk, which are current ratio, quick ratio, inventory turnover ratio, asset-liability ratio, tangible net worth debt ratio, net asset interest rate, and multiples of interest earned. The above indicators are specifically defined as follows:

$x1$ = current ratio = total current assets/total current liabilities. This index reflects the company's ability to repay short-term debt. The more the current assets and the fewer the short-term debts, the greater the current ratio and the stronger the company's short-term debt repayment ability.

$x2$ = quick ratio = (total current assets−inventory)/total current liabilities.

This index can reflect the company's ability to repay short-term debt. Because current assets still include inventories that have a slower realization rate and may have depreciated, the current assets are deducted from inventories and then compared with current liabilities to measure the company's short-term debt solvency.

$x3$ = inventory turnover rate = product sales cost/[(beginning inventory + ending inventory)/2]. This index is the main indicator of inventory turnover speed. Carrying high inventory turnover rate and shortening the business cycle can improve the company's liquidity.

$x4$ = asset – liability ratio = (total liabilities/total assets) ×100%. This index reflects the ratio of capital provided by creditors to total capital. This index is also called the debt-to-business ratio.

$x5$ = tangible net worth debt ratio = [total liabilities/(shareholder equity-net intangible assets)]×100%.

The extension of the property rights ratio index more cautiously and conservatively reflects the degree to which the capital invested by creditors is protected by shareholders' rights during the liquidation of the enterprise. Regardless of the value of intangible assets, including goodwill, trademarks, patent rights, and nonpatent technologies, they may not be used to repay debts. For the sake of caution, they will all be regarded as insolvent.

$x6$ = net asset interest rate = net profit/[(total assets at the beginning of the period + total assets at the end of the period)/2] × 100%. This index compares the net profit of the company for a certain period with the company's assets, showing the comprehensive utilization effect of the company's assets. The higher the index, the higher the efficiency of asset utilization, indicating that the company has achieved good results in increasing income and saving funds. Otherwise, the opposite conclusion is true.

$x7$ = multiple of interest earned = profit before interest and tax/interest expense = (total profit + financial expenses)/(interest expense in financial expenses + capitalized interest).

The ratio of business income to interest expense is used to measure the company's ability to repay the interest on borrowings. It is also called interest protection multiple. As

long as the multiple of the interest earned is large enough, the enterprise has sufficient ability to repay the interest.

3. Models of Risk Assessment

This paper mainly selects three types of machine learning algorithms: RF, SVM, and AdaBoost, to train enterprise risk assessment models. Their basic principles are introduced as follows [18–24].

3.1. RF. RF is one of the most commonly used and most powerful supervised learning algorithms, which takes into account the ability to solve regression and classification problems. Random forest is an algorithm that integrates multiple decision trees through the idea of ensemble learning. For the classification problems, the output category is determined by the mode of individual tree output. In the regression problem, the output of each decision tree is averaged to get the final regression result. The specific steps of the RF algorithm are as follows:

(1) The bootstrap resampling method is applied to randomly sample s subtraining sets with replacement in the original dataset to form s decision trees, namely, D_1, D_2, D_3..., D_s. The s value is selected according to the stability of the error curve of the model.

(2) The number m of preselected variables of the tree node is specified, that is, m variables are randomly generated for the construction of the binary tree on the node. The m value is selected by successively calculating the residual sum of squares of the model, so that the m value with the smallest residual sum of squares is the optimal number of variables.

(3) For a single decision tree, the nodes are recursively partitioned according to the principle of minimum node impurity (that is, the Gini coefficient is the smallest) among the m variables. The Gini coefficient is defined as follows:

$$\text{Gini}(t) = 1 - \sum_{j} [p(j|t)]^2, \tag{1}$$

where t is a decision tree node and $p(j|t)$ is the probability of category j at node t.

(4) Each decision tree is traversed and step (3) is repeated. The decision tree grows arbitrarily without pruning operations.

(5) The s decision trees form a forest, and the voting method is used to determine and classify the classified data.

3.2. SVM. The basic idea of SVM is to map the data to the high-dimensional feature space through nonlinear mapping and realize the linear regression transformation from the nonlinear function estimation problem to the high-dimensional feature space. The training samples are denoted as $(x_i, y_i), i = 1, 2, \cdots, N$, $x_i \in R^n$ is the input vector, $y_i \in R$ is the corresponding output value, and, N is the number of training samples. The linear model of the high-dimensional space can be expressed as follows:

$$f(x, \omega) = \sum_{j=1}^{m} \omega_j \Phi(x)_j + b, \tag{2}$$

where x is the input vector; ω is the feature space coefficient vector; $\Phi(x)_j, j = 1, 2, \cdots, m$, is the nonlinear transfer function; $\omega_j (j = 1, 2, \cdots, m)$ is the coefficient of the corresponding $\Phi(x)_j$ feature space; and b is the deviation term of the high-dimensional space. The structural risk function $R(\omega)$ is constructed as follows:

$$R(\omega) = \frac{1}{2}\omega^2 + C \sum_{i=1}^{N} L_\varepsilon (y_i, f(x_i, \omega)), \tag{3}$$

where $\|\omega\|$ is the Euclidean distance of the feature space coefficient vector; C is the penalty coefficient; and $L_\varepsilon (y_i, f(x_i, \omega))$ is the loss function, in which $y_i (i = 1, 2, \cdots, N)$ is the sample output value and $f(x_i, \omega) (i = 1, 2, \cdots, N)$ is the output value of the corresponding x_i in high-dimensional space.

This paper uses a linear insensitive loss function, which is defined as follows:

$$L_\varepsilon (y_i, f(x_i, \omega)) = \begin{cases} 0, & |f(x, \omega) - y| < \varepsilon, \\ |f(x, \omega) - y| - \varepsilon, & |f(x, \omega) - y| > \varepsilon. \end{cases} \tag{4}$$

In order to minimize the structural risk function $R(\omega)$, the regression equation can be written as

$$\begin{aligned} f(x) &= \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) K(x_i, x) + b, \\ \sum_{i=1}^{N} (\alpha_i - \alpha_i^*) &= 0, \quad \alpha_i, \alpha_i^* \in [0, C], \end{aligned} \tag{5}$$

where α_i and α_i^* are the Lagrangian multipliers, which can be solved by the minimum optimization algorithm of the dual problem sequence, and the kernel function K is defined as the inner product of the eigenvectors after nonlinear transformation, i.e.,

$$K(x_i, x) = \langle \Phi(x_i), \Phi(x) \rangle. \tag{6}$$

Any function that satisfies Mercer's condition can be used as a kernel function. If the kernel function coefficient corresponding to a sampling point is not zero, then the sampling point is a support vector. The commonly used kernel functions in SVM include Gaussian kernel function, radial basis kernel function, etc.

3.3. AdaBoost. This paper is based on single-label multi-class problems, so we choose the simpler and direct AdaBoost algorithm. The main steps of the algorithm are as follows:

(1) The weight distribution of training data points is initialized. The weak learner iteratively operates T

times and produces a weak hypothesis h: $X \longrightarrow Y$ after each iteration. The T value can be selected according to the error curve of the final strong classification.

(2) The calculation of classification error rate is performed using the following formula:

$$h_t : \xi_t = \sum_{i:\, h_t(x_i) \neq y_i} D_t(i), \tag{7}$$

where D_t is the weight distribution of the training data at the tth iteration. In each iteration, if $\xi_t > 1/2$, then this iteration will be aborted.

(3) The weight is assigned to the weak hypothesis according to the classification error rate, and the weight distribution of training data points is updated as follows:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} \beta_t \, h_t(x_i) = y_i \\ 1 \quad \text{others} \end{cases}, \tag{8}$$

where $\beta_t = \xi_t / (1 - \xi_t)$ and Z_t is the normalization constant.

(4) All the weak hypotheses with weights are combined into the final prediction function. The calculation formula is as follows:

$$h_{\text{fin}}(x) = \arg\max_{y \in Y} \sum_{t=1}^{T} \ln\left(\frac{1}{\beta_t}\right) [h_t(x) = y]. \tag{9}$$

The basic idea of the method in this paper is described in Figure 1. Based on the historical training data, the indicator feature vector is constructed according to the method described in Section 2. Accordingly, three types of machine learning algorithms are trained to obtain evaluation models. In the test phase, for the acquired data, the index feature vector is also constructed, and the training evaluation model is input to obtain the current enterprise's risk evaluation result.

4. Experiments and Analysis

4.1. Dataset and Evaluation Indicators. The data sample used in this paper is to select 300 loan companies from a bank and divide them into two categories, i.e., "performance companies (y=1)" and "default companies (y=−1)" according to their financial status, operating status, and past credit records. According to the established safety evaluation index system, each sample is a 7-dimensional vector. First of all, the sample data are processed for robustness and efficiency. In view of the large sample data volume and the smoothness of the data, the double triple standard deviation test is used to eliminate abnormal data, and the total number of effective samples is finally obtained as 500. Among them, 255 companies are able to repay bank credit loans, and the remaining 245 are unable to repay loans on time.

In order to quantitatively analyze the performance of the proposed method, this paper selects accuracy and ROC curve as evaluation indicators. Among them, the accuracy index is a simple and effective index for evaluating classification and prediction performance and refers to the proportion of the correct evaluation samples in the total samples. Area under the curve (AUC) can measure the posterior probability, classification performance, and ranking performance of machine learning algorithms, so it has been widely used in the field of machine learning algorithms. Taking false positive class rate (FPR) as the horizontal axis and true positive class rate (TPR) as the vertical axis, a set of different (FPR, TPR) points can be obtained on the coordinate axis by continuously adjusting the classifier threshold. These points are connected into a line to get the ROC curve of the classifier. The ROC curve cannot be directly used as the evaluation index of the classifier, and the AUC value is generally used as the quantitative criterion.

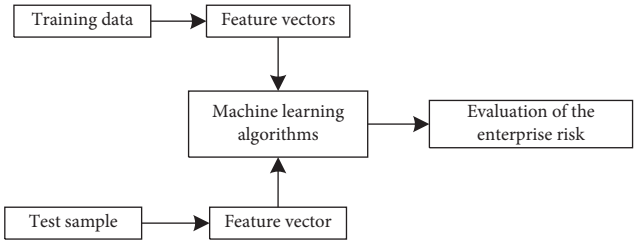

Figure 1: Basic procedure of the proposed method.

5. Result and Analysis

This paper uses K-fold cross validation. Generally, K is 10 because it has relatively low bias and variance. Therefore, this paper divides 500 corporate risk data into 10 equal parts, namely, T1, T2, T3,..., T10. Take Ti as the test dataset, and the remaining part is the training dataset, thereby constructing the ith group of test training sets (Test$i$, Train$i$) (i = 1, 2, ..., 10). The average of the accuracy value and AUC value of each model is calculated, and the statistical results are shown in Table 1.

The following can be seen from Table 1. (1) Combining the two evaluation index standards, the SVM model is effective, and the RF and AdaBoost models have excellent performance. (2) From the perspective of accuracy, the AdaBoost model is better than the SVM and RF models; from the perspective of the AUC value, the RF model is almost the same as the AdaBoost model, and both are better than the SVM model.

Considering the two evaluation indicators, the accuracy value of the AdaBoost model is 1.2% higher than that of the RF model, and the AUC value is higher than that of the RF model. The relationship between enterprise risk levels is slightly better than SVM and RF models.

Taking into account the possible noise impact of actual data, this paper applies different degrees of noise conditions to 500 sample data and uses signal-to-noise ratio (SNR) to measure the noise level. Figure 2 shows the accuracy performance curves of the three methods under different SNRs. It can be seen from the comparison that the noise robustness

Table 1: Comparison of performance of the three machine learning algorithms.

	Accuracy (%)	AUC
RF	87.9	0.861
SVM	85.2	0.837
AdaBoost	90.1	0.878

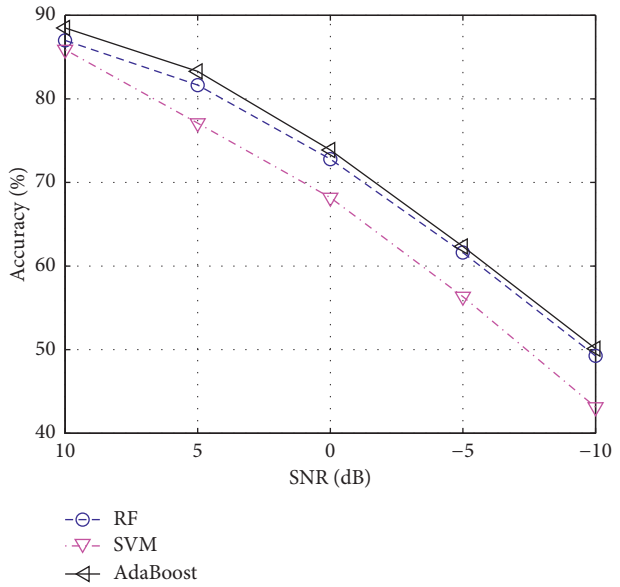

Figure 2: Accuracy of the three machine learning algorithms.

of the RF and AdaBoost methods is still better than that of the SVM method, reflecting its stronger robustness.

6. Conclusion

Statistical learning methods are widely used in risk assessment due to their simple structure and strong interpretation. However, based on the assumption that there is a linear relationship between variables, the prediction effect lacks accuracy and cannot fully reflect the risk status in many cases. The risk assessment model constructed by modern machine learning methods has high accuracy through data training and has broad application prospects in enterprise risk assessment. In this paper, three machine learning algorithms of RF, SVM, and AdaBoost are applied to enterprise risk assessment, which are verified based on actual data. The comparison shows that RF and AdaBoost have higher accuracy in predicting risk. Different machine learning methods have different advantages. Combining different machine learning methods or using integrated learning methods for data feature processing, the performance of the proposed method can be further improved.

Data Availability

The dataset can be accessed upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This study was supported by the Fundamental Research Funds for the Central Universities, Zhongnan University of Economics and Law (2722019JCG070), and Research Project of Humanities and Social Sciences of Hubei Provincial Department of Education (19G009).

References

[1] W. Jiang, L. Deng, L. Chen, J. Wu, and J. Li, "Risk assessment and validation of flood disaster based on fuzzy mathematics," *Progress in Natural Science*, vol. 19, no. 10, pp. 1419–1425, 2009.

[2] a Jian Luo, B. Xin Yan, and Y. Tian, "Unsupervised quadratic surface support vector machine with application to credit risk assessment," *European Journal of Operational Research*, pp. 1008–1017, 2020.

[3] M. Moscatellia, F. Parlapianoa, S. Narizzanob, and G. Viggiano, "Corporate default forecasting with machine learning," *Expert Systems with Applications*, vol. 161, Article ID 113567, 2020.

[4] N. Arora and P. D. Kaur, "A Bolasso based consistent feature selection enabled random forest classification algorithm: an application to credit risk assessment," *Applied Soft Computing Journal*, vol. 86, Article ID 105936, 2020.

[5] D. Paganoti Fonsecaa, P. Fernandes Wankea, and H. Luiz Correa, "A two-stage fuzzy neural approach for credit risk assessment in a Brazilian credit card company," *Applied Soft Computing Journal*, vol. 92, Article ID 106329, 2020.

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] Y. Wanga, Y. Zhanga, L. Yan, and X. Yua, "A comparative assessment of credit risk model based on machine learning—a case study of bank loan data," *Procedia Computer Science*, vol. 174, pp. 141–149, 2020.

[8] J. Jin Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, 2005.

[9] D. D. Wu, D. L. Olson, and C. Luo, "A decision support approach for accounts receivable risk management," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 44, no. 12, pp. 1624–1632, 2014.

[10] J. W. Nowak, S. Sarkani, and T. A. Mazzuchi, "Risk assessment for a national renewable energy target Part II: employing the model," *IEEE Systems Journal*, vol. 10, no. 2, pp. 459–470, 2016.

[11] N. Wan, L. Li, C. Ye, and B. Wang, "Risk assessment in intelligent manufacturing process: a case study of an optical cable automatic arranging robot," *IEEE Access*, vol. 7, pp. 105892–105901, 2019.

[12] Z. Qu, "Application of improved PCA in risks assessment technology of enterprise information security," in *Proceedings of the 2009 2nd International Conference on Power Electronics and Intelligent Transportation System (PEITS)*, pp. 58–61, Shenzhen, China, December 2009.

[13] Z. Xu and G. Chi, "Bank-enterprise project risk assessment model based on the information entropy method," in *Proceedings of the 2011 2nd International Conference on Artificial*

Intelligence, Management Science and Electronic Commerce (AIMSEC), pp. 998–1001, Zhengzhou, China, August 2011.

[14] P.-B. Zhang and Z.-X. Yang, "A novel AdaBoost framework with robust threshold and structural optimization," *IEEE Transactions on Cybernetics*, vol. 48, no. 1, pp. 64–76, 2018.

[15] S. Wu and H. Nagahashi, "Parameterized AdaBoost: introducing a parameter to speed up the training of real AdaBoost," *IEEE Signal Processing Letters*, vol. 21, no. 6, pp. 687–691, 2014.

[16] Qu Qiao, C. Liu, and X. Bao, "E-commerce enterprise supply chain financing risk assessment based on linked data mining and edge computing," *Mobile Information Systems*, vol. 2021, Article ID 9938325, 9 pages, 2021.

[17] T. Chen, Q. Yang, Y. Wang, and S. Wang, "Double-layer network model of bank-enterprise counterparty credit risk contagion," *Complexity*, vol. 2020, Article ID 3690848, 25 pages, 2020.

[18] Y. Shao, X. Yao, L. Tian, and H. Chen, "A multiswarm optimizer for distributed decision making in virtual enterprise risk management," *Discrete Dynamics in Nature and Society*, vol. 2012, Article ID 904815, 24 pages, 2012.

[19] S. Dong and L. Chang, "Sentiment classification for financial texts based on deep learning," *Computational Intelligence and Neuroscience*, vol. 2021, Article ID 9524705, 9 pages, 2021.

[20] X. Wei, "A method of enterprise financial risk analysis and early warning based on decision tree model," *Security and Communication Networks*, vol. 2021, Article ID 6950711, 9 pages, 2021.

[21] J. Wu, C. Li, and Y. Huo, "Safety assessment of dangerous goods transport enterprise based on the relative entropy aggregation in group decision making model," *Computational Intelligence and Neuroscience*, vol. 2014, Article ID 571058, 7 pages, 2014.

[22] A. Galassi and D. José, "Risk assessment of hip fracture based on machine learning," *Applied Bionics and Biomechanics*, vol. 2020, Article ID 8880786, 13 pages, 2020.

[23] B. Williams, B. Allen, Z. Hu et al., "Real-time fall risk assessment using functional reach test," *International Journal of Telemedicine and Applications*, vol. 2017, Article ID 2042974, 8 pages, 2017.

[24] Y. Jianxing, C. Haicheng, W. Shibo, and F. Haizhao, "A novel risk matrix approach based on cloud model for risk assessment under uncertainty," *IEEE Access*, vol. 9, pp. 27884–27896, 2021.

[25] B. R. Greene, S. J. Redmond, and B. Caulfield, "Fall risk assessment through automatic combination of clinical fall risk factors and body-worn sensor data," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 3, pp. 725–731, 2017.