International Journal of Artificial Intelligence in Medical Issues

Volume 3 Issue 2
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v3i2.363

Research Article

Predicting Cardiovascular Disease Using Machine Learning: A Feature Engineering and Model Comparison Approach

Bagus Satrio Waluyo Poetro [1,*]; Dian Hafidh Zulfikar [2]; I Made Sunia Raharja [3]; Nicodemus Mardanus Setiohardjo [4]

[1] Universitas Islam Sultan Agung, Kota Semarang, Jawa Tengah 50112, bagusswp@unissula.ac.id
[2] Universitas Islam Negeri Raden Intan Lampung, Lampung 35131, Indonesia, dianhafidhzulfikar_uin@radenintan.ac.id
[3] Universitas Udayana, Bali 80234, Indonesia, sunia.raharja@unud.ac.id
[4] Politeknik Negeri Kupang, Kota Kupang, Nusa Tenggara Timur. 85258, Indonesia, nicoluck81@gmail.com
Correspondence should be addressed to Bagus Satrio Waluyo Poetro; bagusswp@unissula.ac.id

Received 02 September 2025; Revised 06 September 2024; Accepted 25 October 2024; Published 30 November 2025

Abstract:
Cardiovascular disease (CVD) remains one of the leading causes of mortality globally, emphasizing the need for early detection and effective risk stratification. With the increasing availability of clinical and lifestyle-related health data, machine learning (ML) has become a powerful tool to support data-driven diagnosis and decision-making in healthcare. This study aims to develop and evaluate multiple supervised ML models to predict the presence of cardiovascular disease based on non-invasive features obtained from routine medical checkups. The dataset, comprising 69,301 individual records, includes variables such as age, gender, blood pressure, cholesterol, glucose levels, body measurements, and lifestyle habits. Following comprehensive data cleaning and feature engineering such as the derivation of BMI, Mean Arterial Pressure (MAP), and Pulse Pressure four classifiers were applied: Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM). Model performance was evaluated using metrics including accuracy, precision, recall, F1-score, and ROC-AUC. Among all models tested, the Gradient Boosting Classifier achieved the highest performance, with a ROC-AUC score of 0.8060 and a balanced precision-recall tradeoff, indicating strong discriminatory power. Visualizations such as ROC curves and confusion matrices confirmed the superior capability of Gradient Boosting in differentiating between patients with and without CVD. These findings demonstrate the viability of ML-driven risk assessment models as decision-support tools in clinical settings, potentially aiding in earlier diagnosis and more personalized intervention strategies.

Keywords: Cardiovascular Disease, Machine Learning, Gradient Boosting, Risk Prediction, Health Informatics.

Dataset link: -

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of mortality worldwide, responsible for an estimated 17.9 million deaths annually, accounting for 32% of all global deaths. These conditions, including coronary heart disease, cerebrovascular disease, and other heart-related disorders, are heavily influenced by modifiable risk factors such as hypertension, obesity, smoking, poor diet, and physical inactivity. The global burden of CVD is projected to rise further, particularly in low- and middle-income countries, due to increasing exposure to these risks and limited access to early diagnostics [1], [2], [3].

Traditional risk prediction models, such as the Framingham Risk Score or SCORE, have been widely used for decades. However, these tools often rely on linear assumptions and a limited set of variables, potentially

underestimating individual risks in diverse populations [4], [5]. The emergence of machine learning (ML) presents a transformative opportunity to enhance the prediction of cardiovascular conditions by integrating nonlinear relationships and high-dimensional data from clinical, behavioral, and biometric sources [6], [7]. ML techniques can analyze complex interactions among features such as age, BMI, blood pressure, cholesterol, glucose levels, smoking habits, and physical activity to derive personalized predictions with higher accuracy.

Despite growing interest in ML-based CVD prediction, recent literature highlights several challenges. These include inconsistent feature engineering practices, limited interpretability of black-box models, lack of standardized benchmarks, and the underutilization of longitudinal data [8], [9]. There is also a need for comparative studies that systematically evaluate the performance of various algorithms across large and clean real-world datasets, particularly those that include routine medical check-up indicators [10], [11].

To address this gap, this study proposes a structured ML pipeline to predict cardiovascular disease using a publicly available dataset from routine medical examinations. We introduce advanced feature engineering such as BMI categories, pulse pressure, and composite risk scores and compare the performance of multiple classifiers, including logistic regression, random forest, gradient boosting, and support vector machines. Our approach not only focuses on predictive accuracy but also considers the clinical interpretability of results.

This research is grounded on the following objectives: (1) to evaluate the effectiveness of ML models in predicting cardiovascular disease; (2) to identify the most influential risk factors through engineered features; and (3) to provide empirical evidence supporting data-driven preventive healthcare strategies. By combining robust statistical preprocessing, clear risk stratification, and visual interpretation, this study contributes to the development of transparent, scalable, and efficient ML-based decision support tools in cardiovascular health.

2. Method

This study adopts a machine learning-based pipeline to predict cardiovascular disease (CVD) using patient data derived from clinical examinations. The methodology is divided into six core stages: data acquisition and initial inspection, data cleaning, feature engineering, visualization, model preparation, and evaluation [7], [12], [13]. Each step is discussed in detail as follows.

Figure 1: Research Workflow

Data Loading:

The dataset comprises 69,301 observations with 13 attributes, including demographic, physiological, and lifestyle-related features. The target variable cardio is binary, indicating the presence (1) or absence (0) of cardiovascular disease. As shown in the visualization (see **Figure 2**, top-left), the distribution of the target variable is relatively balanced, with 50.04% of the instances labeled as not having cardiovascular disease and 49.96% labeled as having it. No missing values were found in the dataset, as confirmed during initial inspection.

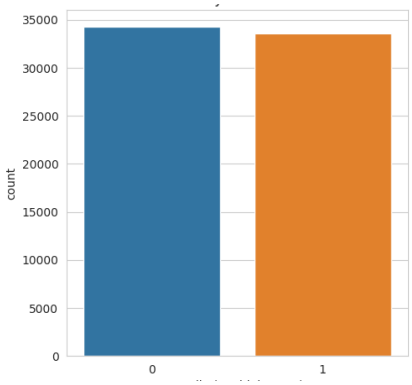

Figure 2: Class Distribution.

Data Cleaning:

To ensure data quality, a cleaning process was conducted by removing biologically implausible values [14]. Blood pressure readings were filtered to retain only values within clinically acceptable ranges (systolic: 80–250 mmHg; diastolic: 50–150 mmHg), and cases where systolic was lower than diastolic were excluded. Records with extreme height (below 130 cm or above 220 cm) or weight (below 30 kg or above 200 kg) were also discarded. Additionally, duplicate records were removed. This process resulted in the exclusion of 1,433 records (2.07%), yielding a cleaned dataset of 67,868 rows.

Feature Engineering:

To enhance predictive performance and incorporate clinical insight, several new features were derived:

- Age in years was computed by converting the original age (in days) using the formula:

$$age_{years} = \frac{age}{365.25} \quad (1)$$

- Body Mass Index (BMI), an established indicator of obesity and cardiovascular risk, was calculated as:

$$BMI = \frac{weight\ (kg)}{\left(height\ (m)\right)^2} \quad (2)$$

- Mean Arterial Pressure (MAP), reflecting the average blood pressure in arteries, was obtained via:

$$MAP = \frac{2 \times ap_{lo} + ap_{hi}}{3} \quad (3)$$

- Pulse Pressure, a marker of arterial stiffness, was derived using:

$$Puls\ Pressure = ap_{hi} - ap_{lo} \quad (4)$$

- Categorical versions of age and BMI were created by binning them into quartile-based risk groups.
- Finally, a composite risk score was created by summing five binary indicators: smoking status (`smoke`), alcohol intake (`alco`), physical inactivity (`active`), elevated cholesterol, and high glucose level.

These engineered features were shown to be important in clinical literature and helped capture nonlinear interactions that raw features alone may not express [15], [16].

Data Preparation:

Before modeling, 15 predictive features were selected, including both original and engineered variables (e.g., `bmi`, `map`, `pulse_pressure`, `risk_score`). The dataset was split into training (80%) and testing (20%) subsets using stratified sampling to preserve the class distribution. All numerical features were standardized using Z-score normalization to ensure comparability across features [17], [18], [19], [20]:

$$Z = \frac{x - \mu}{\sigma} \quad (3)$$

where x is the feature value, μ is the mean, and σ is the standard deviation.

Model Training and Evaluation:

Four supervised machine learning models were trained and evaluated:

- Logistic Regression (LR): A baseline linear classifier that estimates the probability of the target using the logistic (sigmoid) function. It is interpretable and efficient for large datasets [21].
- Random Forest (RF): An ensemble of decision trees that uses bagging and feature randomness to improve generalization and reduce overfitting [22].
- Gradient Boosting Classifier (GBC): An advanced ensemble method that builds trees sequentially, with each new tree correcting the errors of the previous ones. It often yields high performance in classification tasks.
- Support Vector Machine (SVM): A margin-based classifier that attempts to find the hyperplane that best separates the classes. Here, the kernel used allows for non-linear decision boundaries.

Each model was trained using the training data and evaluated using the ROC-AUC score on the test set [23], [24], [25]. Additional performance metrics such as precision, recall, F1-score, and confusion matrix were also computed. The ROC curve comparison (see **Figure 3**, not shown here) visually illustrates the model performance across various thresholds.

3. Result and Discussion

This section presents the outcomes of the machine learning models applied to predict cardiovascular disease, along with a comparative analysis based on multiple evaluation metrics, including precision, recall, F1-score, ROC-AUC, and confusion matrices. The results are also visualized using ROC curves and confusion matrices for each model.

Model Performa Overview:

All four models demonstrated reasonably strong performance in detecting cardiovascular disease. The Gradient Boosting Classifier (GBC) achieved the highest ROC-AUC score of 0.8060, making it the best-performing model overall. It was followed closely by Logistic Regression (0.7960), Support Vector Machine (0.7911), and Random Forest (0.7727). These scores suggest that the models are moderately effective in distinguishing between patients with and without cardiovascular disease.

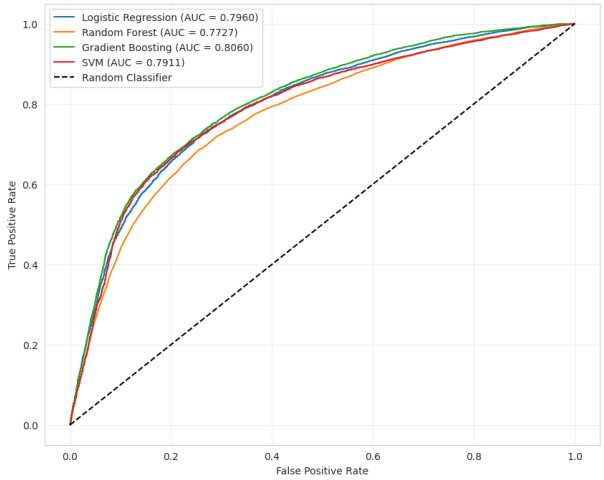

Figure 3: ROC Curve

The ROC curves in **Figure 3** show that Gradient Boosting consistently outperformed the other models across different classification thresholds. It achieved a better balance between the True Positive Rate (Sensitivity) and the False Positive Rate, suggesting superior generalization. The diagonal dashed line represents a random classifier, and all models clearly outperformed it.

Confusion Matrix Interpretation:

Figures 4 to **7** show the confusion matrices for each model:

- Logistic Regression (**Figure 4**): Predicted 4464 out of 6714 actual CVD cases correctly (Recall = 0.66). While its precision for the positive class was 0.76, it also showed a relatively high false positive rate (1426 cases).

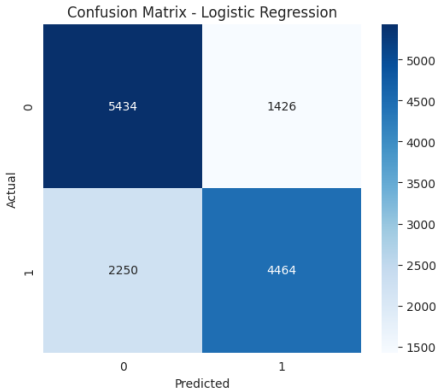

Figure 4: Confusion Matrix of Logistic Regression

- Random Forest (**Figure 5**): Achieved slightly better recall for the positive class (4651/6714 = 0.69) and had fewer false negatives than Logistic Regression, although its overall ROC-AUC was lower.

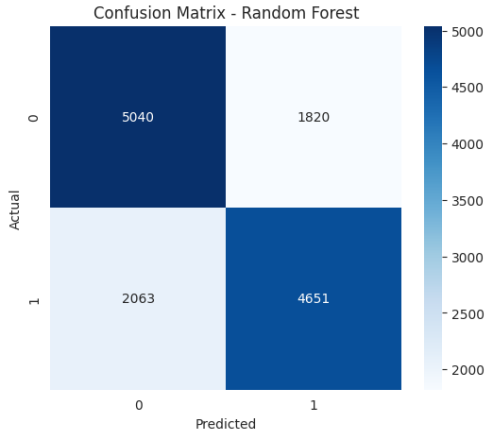

Figure 5: Confusion Matrix of Random Forest

- Gradient Boosting (**Figure 6**): Provided the most balanced predictions, with 4598 true positives and 1473 false positives. Its precision and recall for the positive class were 0.76 and 0.68, respectively, showing a good trade-off between sensitivity and specificity.

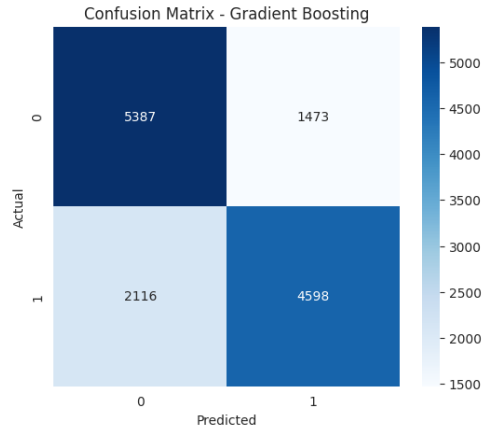

Figure 6: Confusion Matrix of Gradient Boosting

- SVM (**Figure 7**): Yielded strong recall for the negative class (5532/6860 = 0.81) but had a higher number of false negatives (2304), which can be problematic in medical screening where false negatives may delay diagnosis.

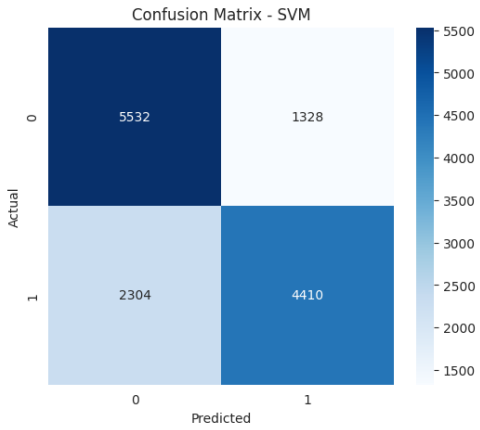

Figure 7: Confusion Matrix of SVM

Table 1. Comparison of Performance Metrics Evaluation

Model	ROC-AUC	Accuracy	Precision (Cardio)	Recall (Cardio)	F1-Score (Cardio)
Logistic Regression	0.796	0.73	0.76	0.66	0.71
Random Forest	0.7727	0.71	0.72	0.69	0.71
Gradient Boosting	**0.806**	**0.74**	**0.76**	**0.68**	**0.72**
SVM	0.7911	0.73	0.77	0.66	0.71

Interpretation and Implications:

The predictive performance of all models reinforces the idea that machine learning can assist clinicians in early CVD detection by leveraging non-invasive data such as age, BMI, blood pressure, and lifestyle habits. The success of Gradient Boosting can be attributed to its sequential learning mechanism, which allows the model to correct misclassifications iteratively, thus enhancing predictive power over simpler models.

Moreover, the engineered features especially Mean Arterial Pressure (MAP), Pulse Pressure, and BMI played a key role in boosting performance, as indicated by the heatmap (**Figure 2**), where these features showed moderate correlation with the target variable cardio.

Interestingly, SVM and Logistic Regression exhibited very similar ROC performance despite their different modeling assumptions. Logistic Regression may be more favorable for clinical implementation due to its interpretability and simplicity, even if its performance is marginally lower than that of GBC.

Overall, the results suggest that Gradient Boosting provides the most robust and reliable prediction performance on this dataset, making it a promising candidate for deployment in clinical decision-support systems.

4. Conclusion

This study successfully demonstrated the application of several machine learning models Logistic Regression, Random Forest, Gradient Boosting, and Support Vector Machine (SVM) for predicting the presence of cardiovascular

disease (CVD) using clinical and behavioral data obtained from routine medical examinations. A systematic pipeline involving data cleaning, feature engineering, scaling, and model evaluation was implemented to ensure data quality and analytical rigor.

Among the models tested, Gradient Boosting Classifier emerged as the best performer, achieving the highest ROC-AUC score of 0.8060 and demonstrating a strong balance between sensitivity and specificity. The model's robustness can be attributed to its iterative nature in learning from misclassified instances, making it particularly effective for complex classification tasks like CVD prediction.

The engineered features such as Body Mass Index (BMI), Mean Arterial Pressure (MAP), and Pulse Pressure provided additional predictive strength and highlighted important physiological patterns associated with cardiovascular risk. Additionally, lifestyle-related variables (smoking, alcohol consumption, and physical activity) were found to have a measurable impact, validating their inclusion in predictive models.

Despite promising results, limitations such as imbalanced recall and precision in some models suggest that further optimization or hybrid ensemble approaches may yield even better performance. Also, while the models are trained on a substantial dataset, external validation on other population groups is necessary to confirm generalizability.

In conclusion, this study underscores the potential of machine learning, particularly Gradient Boosting, in supporting early detection of cardiovascular disease. The integration of such predictive tools into clinical workflows could assist healthcare professionals in identifying high-risk individuals more efficiently, enabling timely intervention and personalized care.

References:

[1] P. P. M. Ramya Sri Bhuvana, B. Rohith, B. M. Swathi, G. Nikhitha, D. H. K. Vege, and D. M. M. Subramanyam, "Prediction Of Cardiovascular Disorders Using Machine Learning," *Educ. Adm. Theory Pract.*, Jun. 2024, doi: 10.53555/kuey.v30i6.5471.

[2] D. Adusumilli, S. L. Damineni, S. Kailasam, N. Tenali, and R. Yadavalli, "Assessment of Cardiovascular Disease Using Machine Learning," *Rev. d'Intelligence Artif.*, vol. 38, no. 3, pp. 1035–1043, Jun. 2024, doi: 10.18280/ria.380329.

[3] M. Jayaraman and S. Pichai, "Automatic Data-Driven Classification Systems for Cardiovascular Disease," *EAI Endorsed Trans. Pervasive Heal. Technol.*, vol. 10, Jun. 2024, doi: 10.4108/eetpht.10.6430.

[4] M. K. S. Bansode, "Heart Disease Prediction using Machine Learning," *Int. J. Res. Appl. Sci. Eng. Technol.*, vol. 12, no. 5, pp. 3124–3128, May 2024, doi: 10.22214/ijraset.2024.62232.

[5] T. Soni, D. Gupta, M. Uppal, and A. Kumari, "Machine Learning in Cardiovascular Disease: Clinical Applications and Relevance to Cardiac Imaging," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, Jul. 2024, pp. 940–944, doi: 10.1109/ICSCSS60660.2024.10624833.

[6] "Predicting Cardiovascular Diseases Risk in Thai Population by Machine Learning," *Bangkok Med. J.*, vol. 21, no. 2, Sep. 2025, doi: 10.31524/bkkmedj.2025.21.006.

[7] D.-I. Kasartzian and T. Tsiampalis, "Transforming Cardiovascular Risk Prediction: A Review of Machine Learning and Artificial Intelligence Innovations," *Life*, vol. 15, no. 1, p. 94, Jan. 2025, doi: 10.3390/life15010094.

[8] H. Sadr, A. Salari, M. T. Ashoobi, and M. Nazari, "Cardiovascular disease diagnosis: a holistic approach using the integration of machine learning and deep learning models," *Eur. J. Med. Res.*, vol. 29, no. 1, p. 455, Sep. 2024, doi: 10.1186/s40001-024-02044-7.

[9] D. Kosaraju, "Machine Learning and the Future of Preventative Cardiology: A Look at Early Detection Techniques," *Galore Int. J. Heal. Sci. Res.*, vol. 8, no. 2, pp. 48–53, Jul. 2024, doi: 10.52403/gijhsr.20230209.

[10] Q. Zheng, "Machine Learning Analysis in the Field of Heart Disease," *Sci. Technol. Eng. Chem. Environ. Prot.*, vol. 1, no. 8, Aug. 2024, doi: 10.61173/qz08vs80.

[11] Raza Naeem, "Machine And Deep Learning Techniques for Cardiovascular Disease Detection," *J. Innov. Comput. Emerg. Technol.*, vol. 4, no. 2, Oct. 2024, doi: 10.56536/jicet.v4i2.131.

[12] M. Begum and D. K. Mahabubullah, "Cardiovascular Disease Prediction Using Machine Learning," *Indian J. Comput. Sci. Technol.*, pp. 360–364, Aug. 2025, doi: 10.59256/indjcst.20250402049.

[13] R. Regen and H. Setiawan, "Advancing Cardiovascular Risk Prediction: A Review of Machine Learning Models and Their Clinical Potential," *J. Electr. Technol. UMY*, vol. 8, no. 2, pp. 51–59, Apr. 2025, doi: 10.18196/jet.v8i2.25208.

[14] A. Mahabub, M. I. Mahmud, and F. Hossain, "A robust system for message filtering using an ensemble machine learning supervised approach," *ICIC Express Lett. Part B Appl.*, vol. 10, no. 9, pp. 805–811, 2019, doi: 10.24507/icicelb.10.09.805.

[15] A. Tuppad and S. D. Patil, "Data Pre-processing Issues in Medical Data Classification," *2023 Int. Conf. ...*, 2023.

[16] G. Ketepalli and P. Bulla, "Data Preparation and Pre-processing of Intrusion Detection Datasets using Machine Learning," *2023 Int. Conf. ...*, 2023.

[17] D. Qi, "Improving Unbalanced Security X-Ray Image Classification Using VGG16 and AlexNet with Z-Score Normalization and Augmentation," *Lecture Notes in Electrical Engineering*, vol. 1182. pp. 205–217, 2024, doi: 10.1007/978-981-97-1463-6_14.

[18] D. Geem, "Progression of Pediatric Crohn's Disease Is Associated With Anti–Tumor Necrosis Factor Timing and Body Mass Index Z-Score Normalization," *Clin. Gastroenterol. Hepatol.*, vol. 22, no. 2, pp. 368–376, 2024, doi: 10.1016/j.cgh.2023.08.042.

[19] M. Sholeh, "Comparison of Z-score, min-max, and no normalization methods using support vector machine algorithm to predict student's timely graduation," *AIP Conference Proceedings*, vol. 3077, no. 1. 2024, doi: 10.1063/5.0202505.

[20] S. Balaji, "Enhancing Diabetic Retinopathy Image Classification using CNN, Resnet, and Googlenet Models with Z-Score Normalization and GLCM Feature Extraction," *Proceedings of the 2nd International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics, ICIITCEE 2024*. 2024, doi: 10.1109/IITCEE59897.2024.10467709.

[21] Akinyemi Moruff Oyelakin and Jimoh Rasheed G, "A Survey of Feature Extraction and Feature Selection Techniques used in Machine Learning-Based Botnet Detection Schemes."

[22] G. V. Titaley, N. Rismayanti, A. N. Handayani, and J. T. Ardiansah, "Performance Comparison of Ensemble Learning Models for Brain Tumor Detection on Augmented MRI Datasets," *Ilk. J. Ilm.*, vol. 17, no. 2, pp. 86–97, Aug. 2025, doi: 10.33096/ilkom.v17i2.2523.86-97.

[23] H. Azis, M. Abdullah, S. Ismail, and ..., "A Comparative Study of YOLO Models for Enhanced Vehicle Detection in Complex Aerial Scenarios," *2025 19th Int. …*, 2025.

[24] Purnawansyah, A. P. Wibawa, and ..., "An in-depth exploration of supervised and semi-supervised learning on face recognition," *Open Computer …*. degruyterbrill.com, 2025, doi: 10.1515/comp-2025-0029.

[25] S. Bharathidason and C. J. Venkataeswaran, "Improving Classification Accuracy based on Random Forest Model with Uncorrelated High Performing Trees," 2014.