Computational Design: Journal of Modern Informatics

Journal of MEDICAL INFORMATICS

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

Predicting Diabetes Mellitus Using Machine Learning

M. Haickal Kamil[1], Hilma Sya'dani Igmu[2], Hafiz Mubaroq[3]

[1,2,3] *State Madrasah Aliyah, West Pasaman Regency, Indonesia*

(Received :July 22, 2026; Revised: July 26, 2026; Accepted: September 10, 2026; Available online: September 30, 2026)

Abstract

Diabetes Mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from defects in insulin secretion, insulin action, or both. The increasing prevalence of diabetes worldwide poses significant challenges to healthcare systems, especially in developing countries. Early detection and prediction are essential to prevent complications and improve patient outcomes. This study aims to develop and evaluate machine learning models for predicting Diabetes Mellitus using clinical datasets. The methodology involves data preprocessing, including missing value handling, normalization, and feature selection, followed by model training using Logistic Regression, Decision Tree, and Random Forest algorithms. The models are evaluated using accuracy, precision, recall, and F1-score metrics. The results indicate that Random Forest provides superior performance compared to other models. This study highlights the potential of machine learning in supporting early diagnosis and decision-making in healthcare systems. Future work includes integrating real-time health data and implementing deep learning approaches for improved prediction accuracy.

Keywords: Diabetes Mellitus, Machine Learning, Prediction, Healthcare Analytics, Random Forest

1. Introduction

Diabetes Mellitus is one of the most prevalent chronic diseases worldwide and is recognized as a major public health concern. According to the World Health Organization, the number of individuals living with diabetes has increased significantly over the past decades, posing a serious burden on healthcare systems. If not properly managed, diabetes can lead to severe complications, including cardiovascular diseases, kidney failure, neuropathy, and retinopathy [1].

Early detection of Diabetes Mellitus is crucial to reduce the risk of complications and to improve patients' quality of life. However, conventional diagnostic approaches often depend on laboratory-based tests, which may not always be accessible or affordable, particularly in low-resource settings. Therefore, there is a growing need for alternative approaches that are efficient, cost-effective, and scalable [2]. Machine learning (ML), a subset of artificial intelligence, has emerged as a powerful tool in healthcare analytics. ML algorithms are capable of analyzing large volumes of data, identifying hidden patterns, and generating predictive models with high accuracy.

In the context of Diabetes Mellitus, ML techniques can be utilized to estimate the likelihood of disease occurrence based on various clinical and demographic factors. This study aims to explore the application of machine learning techniques for predicting Diabetes Mellitus. The main objectives include developing predictive models, comparing their performance, and identifying the most effective algorithm for accurate disease prediction [3].

***Corresponding author: hm0613176@gmail.com**

DOI: https://doi.org/10.35730/a9jsxf06

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

2. LITERATURE REVIEW

2.1. Diabetes Melitus Overview

Diabetes Mellitus (DM) is a chronic metabolic disorder characterized by elevated blood glucose levels (hyperglycemia) resulting from defects in insulin secretion, insulin action, or both. This condition has become a significant global health concern due to its increasing prevalence and its association with severe complications, including cardiovascular disease, kidney failure, neuropathy, and retinopathy [2].

Diabetes Mellitus is generally classified into three main types: Type 1 Diabetes Mellitus, Type 2 Diabetes Mellitus, and gestational diabetes. Type 1 diabetes is an autoimmune condition in which the body's immune system attacks and destroys insulin-producing beta cells in the pancreas, leading to absolute insulin deficiency. It commonly occurs in children and young adults and requires lifelong insulin therapy [3].

Type 2 Diabetes Mellitus (T2DM) is the most prevalent form, accounting for approximately 90–95% of all diabetes cases worldwide. It is primarily caused by insulin resistance, where the body's cells do not respond effectively to insulin, combined with a relative insulin deficiency. T2DM is strongly associated with modifiable lifestyle factors such as obesity, lack of physical activity, unhealthy diet, and smoking. In addition, genetic predisposition, age, and environmental factors also contribute to its development [4].

Gestational diabetes occurs during pregnancy and is characterized by glucose intolerance that is first recognized during gestation. Although it often resolves after childbirth, women with gestational diabetes have a higher risk of developing Type 2 diabetes later in life [5].

The increasing prevalence of Diabetes Mellitus is closely linked to rapid urbanization, changes in dietary patterns, and sedentary lifestyles. According to global health reports, the burden of diabetes continues to rise, particularly in low- and middle-income countries. This trend highlights the importance of early detection, prevention, and effective disease management strategies [6].

Early prediction and intervention play a crucial role in reducing the risk of complications and improving patient outcomes. Advances in technology, particularly in the field of data science and machine learning, have enabled the development of predictive models that can identify individuals at high risk of developing diabetes. These models utilize various clinical and demographic variables such as age, body mass index (BMI), blood pressure, glucose levels, and family history [7].

Furthermore, public health strategies focusing on lifestyle modification—such as promoting healthy eating, regular physical activity, and weight management—are essential in preventing and controlling Type 2 Diabetes Mellitus. Therefore, integrating clinical knowledge with modern analytical approaches is vital to address the growing challenge of diabetes effectively

2.2. Machine Learning in Healthcare

Machine Learning (ML) is a subfield of Artificial Intelligence (AI) that focuses on the development of algorithms capable of learning from data and making predictions or decisions without being explicitly programmed. In the healthcare domain, the application of ML has grown rapidly and is widely used in

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

various areas, including disease diagnosis, risk prediction, medical image analysis, and Clinical Decision Support Systems (CDSS) [8].

One of the key advantages of Machine Learning in healthcare is its ability to process and analyze large volumes of complex data, both structured and unstructured. Structured data includes numerical and categorical information such as age, blood pressure, glucose levels, and laboratory test results. In contrast, unstructured data consists of clinical notes, medical imaging data (such as X-rays, CT scans, and MRIs), and data generated from wearable devices. This capability makes ML particularly suitable for analyzing Electronic Health Records (EHRs), which serve as a primary data source in modern healthcare systems [9].

Machine Learning techniques can generally be categorized into three main types: supervised learning, unsupervised learning, and reinforcement learning. Supervised learning is used when labeled data is available and is commonly applied in classification and prediction tasks, such as diagnosing diseases or predicting patient outcomes. Common algorithms include Logistic Regression, Decision Trees, Random Forest, and Support Vector Machines (SVM). Unsupervised learning, on the other hand, is used to identify hidden patterns or groupings within unlabeled data, such as patient clustering based on similar characteristics. Reinforcement learning is typically applied in dynamic decision-making scenarios, such as optimizing treatment strategies and personalized care planning [10].

The applications of Machine Learning in healthcare are extensive. In disease diagnosis, ML models can assist in detecting conditions more quickly and accurately, including Diabetes Mellitus, cardiovascular diseases, and cancer. In medical image analysis, advanced techniques such as deep learning—particularly Convolutional Neural Networks (CNNs)—are capable of identifying complex patterns in imaging data, thereby supporting radiologists in detecting abnormalities. Additionally, in Clinical Decision Support Systems (CDSS), ML can provide data-driven recommendations to healthcare professionals, including diagnostic suggestions, treatment options, and prognosis predictions [11].

Beyond diagnosis, Machine Learning plays a crucial role in predictive analytics, which involves forecasting the likelihood of diseases before symptoms appear. This is particularly important for chronic diseases such as diabetes, where early detection and prevention can significantly reduce complications. By leveraging historical patient data, ML models can identify high-risk individuals and enable early intervention [12].

Despite its significant advantages, the implementation of Machine Learning in healthcare also presents several challenges. These include issues related to data quality and completeness, patient data privacy and security, and the need for model interpretability (explainability). In clinical settings, it is essential that ML models are not only accurate but also transparent and explainable to gain the trust of healthcare professionals.

In conclusion, Machine Learning holds great potential to improve healthcare delivery through effective data utilization. When properly integrated with clinical expertise and health information systems, ML can support more accurate, efficient, and evidence-based decision-making, ultimately enhancing patient outcomes and the overall performance of healthcare systems.

2.3. Previous Studies

Numerous studies have investigated the application of Machine Learning (ML) techniques for diabetes prediction, aiming to improve early detection and support clinical decision-making. These studies utilize

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

various algorithms and datasets to identify patterns and risk factors associated with the development of Diabetes Mellitus, particularly Type 2 Diabetes.

Among the most commonly used methods, Logistic Regression has been widely applied due to its simplicity, efficiency, and high interpretability. It is particularly suitable for binary classification problems, such as predicting whether a patient is diabetic or non-diabetic. Logistic Regression models provide probabilistic outputs and allow for clear interpretation of the relationship between independent variables (e.g., age, BMI, glucose level) and the likelihood of disease occurrence, making them valuable in clinical settings.

Decision Tree algorithms are also frequently used in diabetes prediction studies. These models generate rule-based classification structures that are easy to visualize and interpret, allowing healthcare professionals to understand the reasoning behind predictions. Decision Trees can handle both numerical and categorical data and are capable of capturing non-linear relationships. However, they are prone to overfitting, especially when the model becomes too complex.

To address the limitations of single decision trees, ensemble learning methods such as Random Forest have been widely adopted. Random Forest combines multiple decision trees to improve predictive performance and robustness. By aggregating the results of several trees, this method reduces overfitting and enhances generalization. Many studies have reported that Random Forest achieves higher accuracy compared to individual models, particularly when dealing with complex and high-dimensional datasets.

In addition to these traditional machine learning techniques, advanced approaches based on deep learning have also gained significant attention. Artificial Neural Networks (ANNs), including Deep Neural Networks (DNNs), are capable of modeling complex and non-linear relationships within large datasets. These models have shown promising results in diabetes prediction, especially when applied to large-scale and high-dimensional data such as Electronic Health Records (EHRs) and continuous monitoring data from wearable devices.

Furthermore, recent studies have explored hybrid and comparative approaches, where multiple algorithms are evaluated to determine the most effective model for diabetes prediction. Techniques such as Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Gradient Boosting methods (e.g., XGBoost) have also demonstrated strong performance in various datasets. These comparative studies highlight that no single algorithm is universally superior; instead, model performance depends on data characteristics, feature selection, and preprocessing techniques.

Despite the promising results, several challenges remain. Deep learning models, while powerful, often require substantial computational resources and large amounts of labeled data. Additionally, they are generally considered “black-box” models due to their limited interpretability, which can hinder their adoption in clinical practice. In contrast, traditional machine learning models, although sometimes less accurate, offer better transparency and are easier to interpret by healthcare professionals.

In conclusion, previous studies demonstrate that Machine Learning techniques have significant potential in improving diabetes prediction. The choice of algorithm should consider not only predictive accuracy but also interpretability, computational efficiency, and applicability in real-world clinical settings. Therefore, ongoing research continues to focus on developing models that balance performance and explainability to support effective healthcare decision-making.

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

2.4. Research Gap

Despite the rapid advancement and widespread application of Machine Learning (ML) techniques in diabetes prediction, several important gaps remain in the existing body of research. Many previous studies have primarily focused on the development and evaluation of individual algorithms, often emphasizing accuracy as the main performance metric. However, such an approach may not provide a comprehensive understanding of model effectiveness, especially in real-world clinical settings where multiple factors must be considered.

One of the key limitations in prior research is the lack of comprehensive comparative studies involving multiple machine learning algorithms applied to the same dataset under consistent conditions. Differences in data preprocessing, feature selection, dataset size, and evaluation methods often make it difficult to directly compare the performance of various models across studies. As a result, it remains unclear which algorithm performs best under specific conditions or for particular types of data.

In addition, many studies rely heavily on a single evaluation metric, such as accuracy, which can be misleading, particularly when dealing with imbalanced datasets a common issue in medical data. For instance, in diabetes prediction, the number of non-diabetic cases may significantly outweigh diabetic cases, leading to biased results if only accuracy is considered. Therefore, other evaluation metrics such as precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC-ROC) are essential to provide a more balanced and reliable assessment of model performance.

Another important gap lies in the limited focus on model interpretability and clinical applicability. While advanced models such as deep learning algorithms often achieve high predictive performance, they are typically less interpretable and function as “black-box” systems. This lack of transparency can reduce trust and hinder adoption among healthcare professionals, who require clear and explainable insights to support clinical decision-making.

Furthermore, there is still a need for studies that integrate both performance optimization and practical implementation aspects, including computational efficiency, scalability, and ease of integration with existing healthcare systems such as Electronic Health Records (EHRs). Many existing studies are conducted in controlled experimental settings and may not fully address the challenges of real-world deployment.

Therefore, this study aims to address these gaps by conducting a comprehensive comparative analysis of multiple machine learning algorithms for diabetes prediction using a consistent dataset and standardized evaluation metrics. By evaluating models based on multiple performance indicators and considering both accuracy and interpretability, this research seeks to identify the most effective and practical approach for supporting early diabetes detection. Ultimately, the findings are expected to contribute to the development of more reliable, transparent, and clinically applicable predictive models in healthcare.

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

3. Methodology

3.1 Research Design

This study adopts a quantitative research approach utilizing secondary data to develop and evaluate machine learning models for diabetes prediction. The research is designed to systematically assess the performance of multiple classification algorithms under consistent conditions.

The overall workflow of this study consists of several sequential stages. First, data collection is performed by obtaining a publicly available dataset relevant to diabetes prediction. Second, data preprocessing is conducted to ensure data quality and suitability for modeling. Third, machine learning models are developed using selected algorithms. Fourth, model evaluation is carried out using standardized performance metrics. Finally, the results are analyzed and compared to identify the most effective model.

This structured approach ensures that the research process is reproducible, objective, and suitable for comparative analysis.

3.2 Dataset Description

The dataset used in this study consists of medical attributes commonly associated with diabetes risk and diagnosis. These features are widely used in clinical and research settings due to their relevance in identifying patterns related to Diabetes Mellitus.The dataset includes the following variables on table 1.

Table 1. Dataset Variables and Description

Variables	Description
Glucose Level	Represents the concentration of glucose in the blood, which is a primary indicator of diabetes.
Blood Pressure	Measures the force of blood against artery walls; abnormal levels are often associated with metabolic disorders.
Skin Thickness	Refers to the thickness of the skin fold, often used as an indirect measure of body fat.
Insulin Level	Indicates the amount of insulin in the blood, which plays a crucial role in glucose regulation.
Body Mass Index (BMI)	A measure of body fat based on height and weight, commonly associated with obesity and diabetes risk.
Age	Represents the age of the individual, as the risk of diabetes increases with age.
Diabetes Pedigree Function	A function that estimates the likelihood of diabetes based on family history and genetic predisposition.

These attributes provide a comprehensive representation of physiological and demographic factors influencing diabetes risk, making the dataset suitable for predictive modeling.

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

3.3 Data Preprocessing

Data preprocessing is a critical step in machine learning to ensure data quality, consistency, and optimal model performance. In this study, several preprocessing techniques were applied:

a. Handling Missing Values

Missing or zero values in the dataset were treated using mean imputation, where missing entries are replaced with the mean value of the respective feature. This approach helps maintain dataset completeness without significantly distorting data distribution.

b. Normalization (Feature Scaling)

Since the dataset contains features with different scales (e.g., glucose vs. age), normalization was performed to standardize the range of values. Techniques such as Min-Max Scaling or Standardization were applied to ensure that all features contribute equally to model training.

c. Feature Selection

Relevant features were selected based on their contribution to the prediction task. This step helps reduce dimensionality, eliminate irrelevant or redundant variables, and improve model efficiency and accuracy.

d. Data Splitting

The dataset was divided into training and testing sets (commonly 80:20 or 70:30 ratio) to evaluate model generalization on unseen data.

e. Outlier Detection (Optional Enhancement)

Outliers were examined and handled where necessary, as extreme values may negatively impact model performance.

3.4 Model Development

To establish a reliable predictive framework in clinical decision-making, selecting appropriate computational models is essential for balancing predictive accuracy with interpretability. Machine learning has emerged as a vital tool in medical diagnostics, offering diverse methodological approaches to analyze complex patient data, identify subtle pathological patterns, and evaluate disease risks. To address these clinical requirements and compare different analytical paradigms—ranging from standard statistical estimation to rule-based modeling and ensemble strategies—three distinct machine learning algorithms were evaluated in this study

a. Logistic Regression

Logistic Regression is a statistical classification technique used for binary outcomes. It estimates the probability that a given input belongs to a particular class using a logistic (sigmoid) function. This model is widely used in healthcare due to its simplicity, efficiency, and high interpretability. It allows researchers and clinicians to understand the relationship between input variables and the probability of disease occurrence.

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

b. Decision Tree

Decision Tree is a supervised learning algorithm that models decisions using a tree-like structure. The dataset is recursively split into subsets based on feature values, forming branches and nodes. Each leaf node represents a classification outcome. Decision Trees are easy to interpret and visualize, making them suitable for explaining decision-making processes in clinical environments. However, they may suffer from overfitting if not properly controlled.

c. Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to produce a final prediction. Each tree is trained on a random subset of the data and features, which increases diversity among the models. This approach improves prediction accuracy, enhances robustness, and reduces overfitting compared to a single decision tree. Random Forest is particularly effective when dealing with complex datasets.

3.5 Model Evaluation Metrics

To evaluate and compare the performance of the machine learning models, several evaluation metrics were used:

a. Accuracy
   Measures the proportion of correctly classified instances among all observations. While widely used, accuracy alone may not be sufficient for imbalanced datasets.

b. Precision
   Indicates the proportion of correctly predicted positive cases out of all predicted positives. It reflects the model's ability to avoid false positives.

c. Recall (Sensitivity)
   Measures the proportion of actual positive cases that are correctly identified. It reflects the model's ability to detect true cases of diabetes.

d. F1-Score
   The harmonic mean of precision and recall, providing a balanced measure of model performance, especially when dealing with imbalanced data.

These evaluation metrics provide a comprehensive assessment of classification performance, enabling a more reliable comparison of different machine learning models.

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

4. Result and Discussions

4.1. Experimental Results

The performance of the machine learning models was evaluated using four standard classification metrics: Accuracy, Precision, Recall, and F1-Score. The results of the experiments are summarized in Table 2.

Table 2. Experimental Model Results

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	78%	75%	72%	73%
Decision Tree	80%	78%	76%	77%
Random Forest	85%	83%	81%	82%

From the results, it is evident that all models achieved satisfactory performance, with Random Forest outperforming the other algorithms across all evaluation metrics. Logistic Regression showed consistent but comparatively lower performance, while Decision Tree achieved moderate results.

4.2. Analysis of Results

The experimental results indicate that the Random Forest model achieved the highest accuracy (85%) along with superior precision (83%), recall (81%), and F1-score (82%). This performance demonstrates the effectiveness of ensemble learning techniques in handling complex and high-dimensional healthcare datasets. By aggregating multiple decision trees, Random Forest reduces variance and improves robustness, leading to better generalization on unseen data.

Logistic Regression, although simpler, demonstrated stable and reliable performance with an accuracy of 78%. Its strength lies in its interpretability and efficiency, making it suitable for applications where transparency is critical. However, its linear nature limits its ability to capture complex and non-linear relationships among features, which may explain its relatively lower performance compared to more advanced models.

The Decision Tree model achieved an accuracy of 80%, outperforming Logistic Regression but still falling behind Random Forest. While Decision Trees are intuitive and easy to interpret, they are prone to overfitting, especially when the tree becomes too deep. This tendency can reduce their ability to generalize well to new data, which may have affected its overall performance.

Overall, the comparative analysis highlights the trade-off between model interpretability and predictive performance. While Random Forest provides the highest accuracy, simpler models such as Logistic Regression offer better explainability.

4.3. Discussion

The findings of this study suggest that ensemble learning methods, particularly Random Forest, provide superior predictive performance and generalization capability compared to single classifiers. By combining

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

multiple decision trees, Random Forest effectively minimizes overfitting and captures complex interactions among features, which are common in medical datasets.

The use of multiple clinical and demographic features—such as glucose level, BMI, insulin level, and age—enables machine learning models to identify intricate patterns associated with diabetes risk. These patterns are often difficult to detect using traditional statistical approaches, highlighting the added value of ML techniques in healthcare analytics.

Furthermore, the results demonstrate the potential of machine learning models to function as Clinical Decision Support Systems (CDSS). By providing data-driven predictions, these models can assist healthcare professionals in early diagnosis, risk assessment, and treatment planning, ultimately improving patient outcomes.

However, several challenges must be addressed before widespread clinical implementation. First, data quality remains a critical issue, as missing, inconsistent, or noisy data can significantly affect model performance. Second, model interpretability is essential in healthcare settings, where clinicians require transparent and explainable results to support decision-making. Third, ethical considerations, including patient data privacy, security, and informed consent, must be carefully managed to ensure compliance with healthcare regulations.

Additionally, the generalizability of the models may be limited if trained on a specific dataset. Therefore, external validation using diverse and real-world datasets is necessary to confirm the robustness of the models.

5. CONCLUSION

This study successfully developed and evaluated multiple machine learning models for predicting Diabetes Mellitus using a structured dataset of clinical and demographic features. Among the evaluated algorithms, Random Forest demonstrated the best overall performance, achieving the highest accuracy, precision, recall, and F1-score.

The results confirm that machine learning techniques can serve as effective tools for early disease prediction and can significantly support decision-making processes in healthcare systems. By leveraging data-driven insights, these models have the potential to enhance diagnostic accuracy and improve patient care outcomes.

Despite these promising results, further improvements are needed to enhance model interpretability, scalability, and real-world applicability. Future research should focus on integrating real-time clinical data, exploring advanced techniques such as deep learning, and developing explainable AI (XAI) models to ensure transparency and trust in clinical environments.

References

1. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, dkk. A guide to deep learning in healthcare. Nat Med. Januari 2019;25(1):24–9. doi:10.1038/s41591-018-0316-z

Vol. 1, No. 2, September 30, 2026, pp. 117 - 127

2. Hosmer DW, Lemeshow S, Sturdivant RX. Applied Logistic Regression [Internet]. 1 ed. Wiley; 2013 [dikutip 5 Mei 2026]. (Wiley Series in Probability and Statistics). Tersedia pada: https://onlinelibrary.wiley.com/doi/book/10.1002/9781118548387 doi:10.1002/9781118548387

3. Kuhn M, Johnson K. Applied Predictive Modeling [Internet]. New York, NY: Springer New York; 2013 [dikutip 5 Mei 2026]. Tersedia pada: http://link.springer.com/10.1007/978-1-4614-6849-3 doi:10.1007/978-1-4614-6849-3

4. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, dkk. Clinical information extraction applications: A literature review. Journal of Biomedical Informatics. Januari 2018;77:34–49. doi:10.1016/j.jbi.2017.11.011

5. Nsengimana F, T. Yombouno A. Comparison of Physical Activity Level between Persons with and Persons without Diabetes in Kenya. AJPHR. 15 Desember 2024;12(4):121–4. doi:10.12691/ajphr-12-4-8

6. Si Y, Du J, Li Z, Jiang X, Miller T, Wang F, dkk. Deep representation learning of patient data from Electronic Health Records (EHR): A systematic review. Journal of Biomedical Informatics. Maret 2021;115:103671. doi:10.1016/j.jbi.2020.103671

7. Magliano D, Boyko EJ. IDF diabetes atlas. 10th edition. Brussels: International Diabetes Federation; 2021. 1 hlm.

8. Rajkomar A, Dean J, Kohane I. Machine Learning in Medicine. N Engl J Med. 4 April 2019;380(14):1347–58. doi:10.1056/NEJMra1814259

9. Kaur H, Kumari V. Predictive modelling and analytics for diabetes using a machine learning approach. ACI. 1 Maret 2022;18(1/2):90–100. doi:10.1016/j.aci.2018.12.004

10. Breiman L. Random Forests. Machine Learning. Oktober 2001;45(1):5–32. doi:10.1023/A:1010933404324

11. The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011 (1).