Comparative Analysis of Voting and Stacking Ensemble Learning for Heart Disease Prediction: A Machine Learning Approach

Gregorius Airlangga[1)*)]

[1)]Information Systems Study Program, Atma Jaya Catholic University of Indonesia
[*)]**Correspondence author**: gregorius.airlangga@atmajaya.ac.id, Jakarta, Indonesia

DOI: https://doi.org/10.37012/jtik.v11i1.2584

Abstract

Heart disease remains a leading cause of mortality worldwide, necessitating the development of accurate predictive models for early diagnosis and intervention. This study investigates the effectiveness of ensemble learning approaches, particularly Voting and Stacking classifiers, in comparison to traditional machine learning models and deep learning architectures. Using a dataset containing clinical and diagnostic attributes, preprocessing steps such as label encoding and standardization were applied to ensure compatibility with machine learning models. The ensemble classifiers were constructed using base learners, including Random Forest, Gradient Boosting, and XGBoost, with soft voting aggregation and logistic regression meta-learning for the Stacking approach. The models were evaluated using stratified ten-fold cross-validation based on precision, recall, F1-score, and ROC-AUC. The results indicate that the Voting classifier achieved the highest overall F1-score (0.8882) and ROC-AUC (0.8697), surpassing the Stacking classifier. Additionally, ensemble models demonstrated competitive performance compared to deep learning architectures, with Random Forest and Gradient Boosting achieving the highest ROC-AUC scores of 0.9313 and 0.9279, respectively. The findings suggest that ensemble methods provide an effective, interpretable, and computationally efficient alternative to deep learning for heart disease prediction. This study highlights the potential of ensemble learning in medical applications and provides valuable insights into optimizing classification models for structured tabular healthcare datasets.

Keywords: *Heart Disease Prediction, Ensemble Learning, Voting Classifier, Stacking Classifier, Machine Learning in Healthcare*

Abstrak

Penyakit jantung tetap menjadi penyebab utama kematian di seluruh dunia, sehingga memerlukan pengembangan model prediktif yang akurat untuk diagnosis dan intervensi dini. Studi ini menyelidiki efektivitas pendekatan pembelajaran ensemble, khususnya pengklasifikasi Voting dan Stacking, dibandingkan dengan model pembelajaran mesin tradisional dan arsitektur pembelajaran mendalam. Dengan menggunakan kumpulan data yang berisi atribut klinis dan diagnostik, langkah-langkah praproses seperti pengodean label dan standardisasi diterapkan untuk memastikan kompatibilitas dengan model pembelajaran mesin. Pengklasifikasi ensemble dibangun menggunakan pembelajar dasar, termasuk Random Forest, Gradient Boosting, dan XGBoost, dengan agregasi pemungutan suara lunak dan meta-pembelajaran regresi logistik untuk pendekatan Stacking. Model-model tersebut dievaluasi menggunakan validasi silang sepuluh kali lipat yang berstrata berdasarkan presisi, ingatan, skor F1, dan ROC-AUC. Hasilnya menunjukkan bahwa pengklasifikasi Voting mencapai skor F1 keseluruhan tertinggi (0,8882) dan ROC-AUC (0,8697), melampaui pengklasifikasi Stacking. Selain itu, model ensemble menunjukkan kinerja yang kompetitif dibandingkan dengan arsitektur deep learning, dengan Random Forest dan Gradient Boosting yang masing-masing mencapai skor ROC-AUC tertinggi, yaitu 0,9313 dan 0,9279. Temuan tersebut menunjukkan bahwa metode ensemble memberikan alternatif yang efektif, dapat ditafsirkan, dan efisien secara komputasi untuk deep learning dalam prediksi penyakit jantung. Studi ini menyoroti potensi pembelajaran ensemble dalam aplikasi medis dan memberikan wawasan berharga dalam mengoptimalkan model klasifikasi untuk kumpulan data perawatan kesehatan tabular terstruktur.

Kata Kunci: Prediksi Penyakit Jantung, Pembelajaran Ensemble, Pengklasifikasi Pemungutan Suara, Pengklasifikasi Susunan, Pembelajaran Mesin dalam Layanan Kesehatan

INTRODUCTION

Cardiovascular diseases, including heart disease, remain the leading cause of death worldwide, accounting for nearly 17.9 million deaths annually according to the World Health Organization (Gaziano, 2022; Otumo & Asanga, n.d.; Parato, Parato, Fedacko, & Magomedova, 2024). Early detection and timely medical intervention are critical in reducing mortality rates and improving patient outcomes (Barrios, 2022; Gandhi et al., 2023; Organization, 2023). Conventional diagnostic methods, including electrocardiograms, stress tests, and clinical assessments, rely heavily on expert interpretation, which may be subject to human error and variability in diagnosis (Boldireva, 2023). The increasing volume of medical data and the complexity of patient records necessitate automated solutions that can enhance diagnostic accuracy and provide reliable predictions (Tayefi et al., 2021). Machine learning and deep learning models have demonstrated significant potential in analyzing large-scale medical data, identifying patterns, and improving predictive performance compared to traditional rule-based clinical assessments (Lee et al., 2022; Navin, Krishnan, & others, 2024; Papadopoulos, Soflano, Chaudy, Adejo, & Connolly, 2022). However, individual machine learning models often face challenges related to generalizability due to overfitting, data imbalance, and biases in training data (Mathrani, Susnjak, Ramaswami, & Barczak, 2021; Siddique et al., 2023; Tasci, Zhuge, Camphausen, & Krauze, 2022). To overcome these limitations, ensemble learning techniques such as Voting and Stacking classifiers have emerged as effective strategies to improve model robustness and classification accuracy(Rane, Choudhary, & Rane, 2024).

Despite advancements in machine learning and deep learning for healthcare applications, several challenges persist in developing optimal predictive models for heart disease diagnosis (Bhatt, Patel, Ghetia, & Mazzeo, 2023). Model variability is a common issue, as different machine learning algorithms yield varying levels of accuracy, making it difficult to determine the best-performing model (Hafsa, Rushd, & Yousuf, 2023). Data imbalance presents another challenge, as heart disease datasets often contain disproportionate distributions between positive and negative cases, leading to biased predictions that favor the majority class (Al-Alshaikh et al., 2024). Additionally, the complexity of clinical data, which includes both categorical and numerical attributes, necessitates sophisticated feature engineering and preprocessing techniques to enhance model performance (Al-Jamimi, 2024). Another major concern is the generalizability of machine learning models, as they may perform well on training data but fail to maintain the same level of accuracy when exposed to unseen patient records (Rasheed et al., 2022). Addressing these challenges requires robust methodologies that can improve classification accuracy while ensuring reliability across different datasets (Gong, Liu, Xue, Li, & Meng, 2023; Gowal et al., 2021; Subbaswamy, Adams, & Saria, 2021).

Previous studies have extensively explored the use of machine learning models for heart disease prediction (Ahsan & Siddique, 2022; Katarya & Meena, 2021; Ogunpola, Saeed, Basurra, Albarrak, & Qasem, 2024). Traditional classification algorithms such as Decision Trees, Support Vector Machines, Logistic Regression, and Naïve Bayes have been widely used in early research due to their interpretability and ease of implementation (Costa & Pedreira, 2023). However, these models often struggle with handling complex feature relationships in clinical datasets. More recent approaches have leveraged advanced ensemble learning techniques such as Random Forest, Gradient Boosting, and XGBoost, which provide better performance by aggregating multiple weak learners (Demir & Sahin, 2023; Kavzoglu & Teke, 2022; Natras, Soja, & Schmidt, 2022). Deep learning models, including Convolutional Neural Networks and Recurrent Neural Networks, have also been applied to heart disease classification tasks, demonstrating superior feature extraction capabilities in medical imaging and sequential patient data (Behrad & Abadeh, 2022). However, deep learning models typically require large-scale labeled datasets for optimal performance, limiting their effectiveness when applied to structured tabular data with limited samples (Hu et al., 2021). The need for robust and scalable predictive models has led to the increasing adoption of ensemble methods, which combine multiple base classifiers to improve overall performance (Abimannan et al., 2023). Among the most widely used ensemble learning techniques, Voting and Stacking classifiers have gained prominence in medical classification tasks (Mahajan, Uddin, Hajati, & Moni, 2023). The Voting classifier aggregates predictions from multiple models and determines the final outcome using either majority voting or probability-based averaging, while the Stacking classifier integrates multiple base models and employs a meta-learner to refine final predictions (Damaševičius, Venčkauskas, Toldinas, & Grigali\=unas, 2021). Although these methods have shown promising results, a direct comparative analysis between these ensemble strategies and deep learning-based models for heart disease classification remains limited in the existing literature (Zhou et al., 2024).

This study aims to provide a comprehensive comparative analysis of Voting and Stacking ensemble learning strategies for heart disease prediction. The primary objectives are to evaluate the performance of these ensemble-based techniques in comparison with a deep learning model trained with a focal loss function, to address class imbalance issues through appropriate preprocessing and loss function design, to apply stratified ten-fold cross-validation to ensure an unbiased evaluation, and to analyze the trade-offs between computational efficiency, predictive accuracy, and generalization ability. The contribution of this study lies in its structured evaluation of Voting and Stacking classifiers in the context of heart disease prediction, offering valuable insights into the strengths and limitations of ensemble learning for medical diagnosis. By systematically comparing these ensemble techniques with deep learning models, this research provides an in-depth understanding of their effectiveness in handling complex clinical datasets and improving predictive performance.

The rest of this paper is structured as follows. The methodology section describes the dataset, preprocessing steps, and modeling techniques, including the ensemble-based Voting and Stacking classifiers as well as the deep learning model with a focal loss function. The results and discussion section presents the findings obtained through stratified ten-fold cross-validation and compares the performance of each model. The conclusion summarizes key insights, implications, and future directions for enhancing machine learning models in heart disease prediction. By conducting a systematic comparison of ensemble-based models and deep learning strategies, this research aims to contribute to the advancement of artificial intelligence-driven healthcare solutions and provide practical insights for clinical decision support systems.

METHOD

The methodology employed in this study is structured into several key components, including data preprocessing, feature transformation, model formulation, training procedures, and performance evaluation. The primary objective is to systematically compare ensemble learning strategies, specifically Voting and Stacking classifiers, with a deep learning model trained using a focal loss function for heart disease prediction. To ensure a rigorous and unbiased evaluation, stratified ten-fold cross-validation is applied to all models, ensuring robustness across different training and testing splits.

The dataset used in this study consists of (n) patient records with (m) clinical and diagnostic attributes and can be downloaded from (fedesoriano, 2021). Let the dataset be represented as ($\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$), where each feature vector ($x_i \in R^m$) contains patient-specific information and the target label ($y_i \in \{0,1\}$) indicates the presence or absence of heart disease. The features include both categorical and numerical variables, necessitating appropriate preprocessing to ensure compatibility with machine learning models. The first step in preprocessing involves encoding categorical variables. Given a categorical attribute (x_j), label encoding is applied to transform it into an integer representation, such that

$$\tilde{x_j} = f(x_j), \quad f: x_j \rightarrow Z$$

For numerical features, standardization is performed to normalize the distributions. Given a numerical feature (x_j), its standardized value (z_j) is computed as

$$z_j = \frac{x_j - \mu_j}{\sigma_j}$$

where (μ_j) is the mean and (σ_j) is the standard deviation of the feature across all instances. This ensures that all numerical attributes are centered around zero with unit variance, improving model convergence and performance. To enhance predictive accuracy, three models are developed: a Voting classifier, a Stacking classifier, and a deep neural network. The Voting classifier aggregates the predictions of multiple base classifiers by averaging their predicted probabilities. Given a set of base models (h_k), the final prediction $(\hat{y})$ in soft voting is given by

$$\hat{y} = \frac{1}{K}\sum_{k=1}^{K} P\big(y = 1\big|h_k(x)\big)$$

where (K) is the number of base classifiers and $(P\big(y = 1\big|h_k(x)\big))$ represents the probability assigned to the positive class by the (k)-th model. The Stacking classifier refines predictions by introducing a meta-learner. First, predictions from the base models form an intermediate feature matrix

$$H = \begin{bmatrix} h_1(x_1) & h_2(x_1) & \dots & h_K(x_1) \\ h_1(x_2) & h_2(x_2) & \dots & h_K(x_2) \\ \vdots & \vdots & \ddots & \vdots \\ h_1(x_n) & h_2(x_n) & \dots & h_K(x_n) \end{bmatrix} \in R^{n\times K}$$

The meta-learner, denoted as \(g \), then maps \(\mathbf{H} \) to the final prediction:

$$\hat{y} = g(H).$$

For this study, logistic regression is used as the meta-learner due to its efficiency and generalization capability. The deep learning model is a feedforward neural network with multiple hidden layers. Let $(x \in R^m)$ be the input feature vector. The transformation at each layer (l) is given by

$$h^{(l)} = \sigma\big(W^{(l)}h^{(l-1)} + b^{(l)}\big)$$

where $(W^{(l)})$ and $(b^{(l)})$ are the weight matrix and bias vector, respectively, and $(\sigma(\cdot))$ is the ReLU activation function defined as

$$\sigma(x) = \max(0, x)$$

Dropout regularization is applied to mitigate overfitting, ensuring that neurons are randomly deactivated during training with probability (p), leading to a modified activation function

$$h^{(l)} = \text{Dropout}\big(\sigma\big(W^{(l)}h^{(l-1)} + b^{(l)}\big), p\big).$$

The final output layer consists of a single neuron with a sigmoid activation function

$$P(y = 1|x) = \frac{1}{1 + e^{-z}}$$

where (z) is the logit output. The neural network is optimized using the Adam optimizer and trained with a focal loss function to address class imbalance. The focal loss is defined as

$$\mathcal{L}_{focal} = -\alpha(1 - p)^{\gamma} \log(p)$$

where (α) is a weighting factor, (γ) is the focusing parameter, and ($p = P(y = 1|x)$) is the predicted probability. This function reduces the contribution of well-classified examples and emphasizes misclassified instances, improving learning on underrepresented classes. Model performance is evaluated using stratified ten-fold cross-validation. The dataset is divided into ten subsets, with each serving as a test set once while the remaining nine subsets are used for training. This ensures that all instances contribute to both training and testing phases, reducing the variance in performance estimation. The models are assessed based on four key performance metrics. Precision measures the proportion of correctly classified positive cases and is computed as

$$\text{Precision} = \frac{TP}{TP + FP}$$

where (TP) denotes true positives and (FP) denotes false positives. Recall quantifies the ability to identify positive cases and is given by

$$\text{Recall} = \frac{TP}{TP + FN}$$

where (FN) represents false negatives. The F1-score, which balances precision and recall, is defined as

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The ROC-AUC score measures the ability of the model to distinguish between positive and negative cases and is computed as the area under the receiver operating characteristic curve.

RESULTS AND DISCUSSION

The evaluation of the proposed ensemble learning strategies, specifically the Voting and Stacking classifiers, was conducted using stratified ten-fold cross-validation. The performance of these models was compared with other traditional machine learning models, deep learning architectures, and hybrid techniques. The key performance metrics considered include Precision, Recall, F1-score, and ROC-AUC to comprehensively assess classification effectiveness. The results obtained from the experiments highlight the strengths and limitations of different approaches, providing valuable insights into their comparative effectiveness for heart disease prediction.

The Stacking classifier achieved a Precision of 0.8598, Recall of 0.9115, F1-score of 0.8839, and ROC-AUC of 0.8619, while the Voting classifier slightly outperformed it with a Precision of 0.8712, Recall of 0.9076, F1-score of 0.8882, and ROC-AUC of 0.8697. These results indicate that ensemble learning methods effectively combine the strengths of multiple base classifiers to improve classification performance. The Voting classifier, which utilizes a soft-voting strategy, demonstrated better overall performance than Stacking in terms of F1-score and ROC-AUC, suggesting that averaging predictions from diverse models results in a more stable and generalized classifier.

In comparison to individual machine learning models, Random Forest (Precision: 0.8680, Recall: 0.9016, F1-score: 0.8832, ROC-AUC: 0.9313), Gradient Boosting (Precision: 0.8722, Recall: 0.8996, F1-score: 0.8851, ROC-AUC: 0.9279), and XGBoost (Precision: 0.8650, Recall: 0.8997, F1-score: 0.8813, ROC-AUC: 0.9228) delivered strong predictive performance. Notably, the Voting and Stacking classifiers achieved comparable results, demonstrating that ensemble methods successfully enhance prediction accuracy by leveraging multiple models, particularly in complex classification problems such as heart disease detection.

A crucial aspect of this study involves comparing ensemble learning strategies with deep learning models. The Neural Network model, trained using Binary Crossentropy loss, achieved Precision of 0.8756, Recall of 0.8782, F1-score of 0.8758, and ROC-AUC of 0.9232, while the same model trained with Focal Loss performed slightly better in Recall (0.8860) but slightly lower in Precision (0.8697), leading to an F1-score of 0.8763. The deep learning models, including Multi-Layer Perceptron (MLP) and Convolutional Neural Networks (CNNs), demonstrated competitive performance, particularly in terms of ROC-AUC, where they outperformed the ensemble learning approaches. The MLP model with Binary Crossentropy achieved the highest ROC-AUC of 0.9232, followed by MLP with Focal Loss (0.9221), XGBoost (0.9228), and Gradient Boosting (0.9279). This suggests that while deep learning models are capable of capturing complex patterns in the data, they require careful hyperparameter tuning and sufficient computational resources for optimal performance. A comparison between CNN models trained using Focal Loss and Binary Crossentropy further highlights the trade-offs in loss function selection. The CNN model trained with Focal Loss achieved an F1-score of 0.8770 and ROC-AUC of 0.9174, whereas the CNN trained with Binary Crossentropy underperformed in both F1-score (0.8575) and ROC-AUC (0.9051). This suggests that the Focal Loss function is beneficial in handling class imbalance, leading to a better recall rate, which is crucial in medical diagnosis applications.

The weakest performing model in the study was TabNet, which yielded significantly lower performance across all metrics (Precision: 0.4817, Recall: 0.1046, F1-score: 0.1684, and ROC-AUC: 0.5408). This indicates that TabNet may not be well-suited for structured tabular medical datasets with categorical and numerical features unless extensive hyperparameter tuning and feature engineering are performed. The results demonstrate that ensemble learning techniques, particularly the Voting classifier, provide a strong balance between interpretability, computational efficiency, and predictive performance. Unlike deep learning models, which require large-scale data and fine-tuning of hyperparameters, Voting and Stacking classifiers achieve robust classification results without excessive computational costs. The Voting classifier, in particular, achieved the highest F1-score

(0.8882) among all ensemble methods, indicating that its ability to leverage multiple classifiers enhances generalization and stability in heart disease classification tasks.

An important observation from the results is that Stacking classifiers slightly underperformed compared to Voting classifiers, despite being expected to outperform individual base learners. One possible reason is that the meta-learner in the Stacking approach, which is logistic regression, may have limited capacity to capture intricate patterns learned by the base classifiers. The Voting classifier, benefiting from soft voting, allows for a more direct and probabilistic aggregation of predictions, which likely led to its improved performance. Another key takeaway from this study is that machine learning ensembles remain competitive with deep learning models. The best ROC-AUC score was observed in Random Forest (0.9313), followed by Gradient Boosting (0.9279), Support Vector Machine (0.9264), and XGBoost (0.9228), all of which outperformed deep learning models in this study. This result reinforces the notion that ensemble methods such as Random Forest and Gradient Boosting remain highly effective in structured tabular data classification tasks, particularly in medical datasets where feature interactions play a crucial role.

This study primarily focuses on evaluating the effectiveness of Voting and Stacking classifiers in comparison with other machine learning and deep learning approaches for heart disease prediction. The results highlight that ensemble methods offer a compelling alternative to deep learning models, particularly in situations where computational efficiency and interpretability are critical factors. The Voting classifier demonstrated the best overall performance among ensemble methods, showcasing its strength in aggregating diverse classifiers and mitigating weaknesses inherent in individual models. The Stacking classifier, while still competitive, did not significantly outperform its Voting counterpart, suggesting that the selection of base learners and the meta-learner plays a critical role in maximizing Stacking’s advantages.

In contrast, deep learning models, particularly MLP and CNN, exhibited strong performance in ROC-AUC but required extensive fine-tuning to reach optimal results. The use of Focal Loss in deep learning models led to improved Recall but at the cost of slightly reduced Precision, suggesting that it is beneficial in scenarios where identifying positive cases is more critical than minimizing false positives. The results of this study reinforce that ensemble learning approaches, particularly Voting and Stacking, are well-suited for heart disease prediction in structured tabular datasets. The Voting classifier emerged as the most robust model, demonstrating that a soft-voting ensemble can effectively combine the predictive power of multiple base models to achieve superior classification performance. While deep learning methods showed competitive results, they did not significantly outperform ensemble techniques, highlighting the continued relevance of machine learning ensembles in medical classification tasks.

CONCLUSIONS AND RECOMMENDATIONS

This study presented a comparative analysis of Voting and Stacking ensemble learning strategies for heart disease prediction, evaluating their performance against various machine learning and deep learning models. The research aimed to determine the effectiveness of ensemble techniques in enhancing classification accuracy, generalization, and interpretability. Through rigorous stratified ten-fold cross-validation, the models were assessed based on Precision, Recall, F1-score, and ROC-AUC, ensuring reliable and unbiased performance evaluation.

The experimental results indicate that Voting and Stacking classifiers outperformed many traditional machine learning models while providing a computationally efficient alternative to deep learning techniques. The Voting classifier achieved the highest F1-score (0.8882) among ensemble methods, demonstrating its ability to aggregate predictions effectively and enhance classification stability. The Stacking classifier performed competitively, achieving an F1-score of 0.8839, but did not significantly surpass the Voting approach, suggesting that the choice of meta-learner influences the stacking strategy's performance.

Deep learning models, including MLP and CNN trained with both Binary Crossentropy and Focal Loss, exhibited strong ROC-AUC scores, with MLP using Binary Crossentropy achieving the highest ROC-AUC of 0.9232. However, traditional ensemble methods such as Random Forest (ROC-AUC: 0.9313) and Gradient Boosting (ROC-AUC: 0.9279) demonstrated superior classification performance, reaffirming the effectiveness of tree-based ensembles in structured tabular data. Notably, TabNet significantly underperformed, highlighting its limitations in handling medical datasets with categorical and numerical attributes.

A key takeaway from this study is that ensemble learning techniques remain a powerful approach for structured healthcare data classification. The Voting classifier, leveraging soft-voting aggregation, demonstrated its robustness in aggregating multiple base classifiers to achieve superior predictive performance, making it a suitable candidate for practical implementation in clinical decision support systems. The Stacking classifier, while competitive, was limited by the choice of meta-learner, suggesting that future research should explore more complex meta-models to improve performance further. The results also reinforce the continued relevance of ensemble learning in medical AI applications, demonstrating that ensemble techniques can achieve performance levels comparable to deep learning models while maintaining better interpretability and computational efficiency. While deep learning approaches hold promise, their advantages are more pronounced in larger datasets or when dealing with unstructured data such as medical imaging. For structured datasets like clinical records, ensemble learning methods, particularly the Voting classifier, remain a reliable and effective choice.

Future work should explore hybrid models integrating ensemble learning with deep learning architectures, leveraging the strengths of both methodologies. Additionally, further optimization of the Stacking classifier's meta-learning component may enhance its ability to refine predictions, potentially surpassing other ensemble methods. Extending this study to larger and more diverse datasets, incorporating real-world clinical environments, and exploring feature selection techniques could further improve model robustness and generalizability.

REFERENSI

Abimannan, S., El-Alfy, E.-S. M., Chang, Y.-S., Hussain, S., Shukla, S., & Satheesh, D. (2023). Ensemble Multifeatured Deep Learning Models And Applications: A Survey. *IEEE Access*.

Ahsan, M. M., & Siddique, Z. (2022). Machine Learning-Based Heart Disease Diagnosis: A Systematic Literature Review. *Artificial Intelligence In Medicine*, *128*, 102289.

Al-Alshaikh, H. A., P, P., Poonia, R. C., Saudagar, A. K. J., Yadav, M., Alsagri, H. S., & Alsanad, A. A. (2024). Comprehensive Evaluation And Performance Analysis Of Machine Learning In Heart Disease Prediction. *Scientific Reports*, *14*(1), 7819.

Al-Jamimi, H. A. (2024). Synergistic Feature Engineering And Ensemble Learning For Early Chronic Disease Prediction. *IEEE Access*.

Barrios, C. H. (2022). Global Challenges In Breast Cancer Detection And Treatment. *The Breast*, *62*, S3--S6.

Behrad, F., & Abadeh, M. S. (2022). An Overview Of Deep Learning Methods For Multimodal Medical Data Mining. *Expert Systems With Applications*, *200*, 117006.

Bhatt, C. M., Patel, P., Ghetia, T., & Mazzeo, P. L. (2023). Effective Heart Disease Prediction Using Machine Learning Techniques. *Algorithms*, *16*(2), 88.

Boldireva, A. (2023). *Identifying Needs In The Field Of Electrocardiogram Analysis To Increase The Accuracy Of ECG Interpretation*. NTNU.

Costa, V. G., & Pedreira, C. E. (2023). Recent Advances In Decision Trees: An Updated Survey. *Artificial Intelligence Review*, *56*(5), 4765–4800.

Damaševičius, R., Venčkauskas, A., Toldinas, J., & Grigali\=Unas, Š. (2021). Ensemble-Based Classification Using Neural Networks And Machine Learning Models For Windows Pe Malware Detection. *Electronics*, *10*(4), 485.

Demir, S., & Sahin, E. K. (2023). An Investigation Of Feature Selection Methods For Soil Liquefaction Prediction Based On Tree-Based Ensemble Algorithms Using Adaboost, Gradient Boosting, And Xgboost. *Neural Computing And Applications*, *35*(4), 3173–3190.

Fedesoriano. (2021). *Heart Failure Prediction Dataset*. Retrieved From Https://Www.Kaggle.Com/Datasets/Fedesoriano/Heart-Failure-Prediction/Data

Gandhi, Z., Gurram, P., Amgai, B., Lekkala, S. P., Lokhandwala, A., Manne, S., … Others. (2023). Artificial Intelligence And Lung Cancer: Impact On Improving Patient Outcomes. *Cancers*, *15*(21), 5236.

Gaziano, T. A. (2022). Cardiovascular Diseases Worldwide. *Public Health Approach Cardiovasc. Dis. Prev. Manag*, *1*, 8–18.

Gong, Y., Liu, G., Xue, Y., Li, R., & Meng, L. (2023). A Survey On Dataset Quality In Machine Learning. *Information And Software Technology*, *162*, 107268.

Gowal, S., Rebuffi, S.-A., Wiles, O., Stimberg, F., Calian, D. A., & Mann, T. A. (2021). Improving Robustness Using Generated Data. *Advances In Neural Information Processing Systems*, *34*, 4218–4233.

Hafsa, N., Rushd, S., & Yousuf, H. (2023). Comparative Performance Of Machine-Learning And Deep-Learning Algorithms In Predicting Gas--Liquid Flow Regimes. *Processes*, *11*(1), 177.

Hu, W., Fey, M., Ren, H., Nakata, M., Dong, Y., & Leskovec, J. (2021). Ogb-Lsc: A Large-Scale Challenge For Machine Learning On Graphs. *Arxiv Preprint Arxiv:2103.09430*.

Katarya, R., & Meena, S. K. (2021). Machine Learning Techniques For Heart Disease Prediction: A Comparative Study And Analysis. *Health And Technology*, *11*(1), 87–97.

Kavzoglu, T., & Teke, A. (2022). Predictive Performances Of Ensemble Machine Learning Algorithms In Landslide Susceptibility Mapping Using Random Forest, Extreme Gradient Boosting (Xgboost) And Natural Gradient Boosting (Ngboost). *Arabian Journal For Science And Engineering*, *47*(6), 7367–7385.

Lee, S., Shin, J., Kim, H. S., Lee, M. J., Yoon, J. M., Lee, S., … Lee, S. (2022). Hybrid Method Incorporating A Rule-Based Approach And Deep Learning For Prescription Error Prediction. *Drug Safety*, *45*(1), 27–35.

Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning For Disease Prediction: A Review. *Healthcare*, *11*(12), 1808.

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives On The Challenges Of Generalizability, Transparency And Ethics In Predictive Learning Analytics. *Computers And Education Open*, *2*, 100060.

Natras, R., Soja, B., & Schmidt, M. (2022). Ensemble Machine Learning Of Random Forest, Adaboost And Xgboost For Vertical Total Electron Content Forecasting. *Remote Sensing*, *14*(15), 3547.

Navin, K., Krishnan, M., & Others. (2024). Fuzzy Rule Based Classifier Model For Evidence Based Clinical Decision Support Systems. *Intelligent Systems With Applications*, *22*, 200393.

Ogunpola, A., Saeed, F., Basurra, S., Albarrak, A. M., & Qasem, S. N. (2024). Machine Learning-Based Predictive Models For Detection Of Cardiovascular Diseases. *Diagnostics*, *14*(2), 144.

Organization, W. H. (2023). *Global Breast Cancer Initiative Implementation Framework: Assessing, Strengthening And Scaling-Up Of Services For The Early Detection And Management Of Breast Cancer*. World Health Organization.

Otumo, E., & Asanga, D. E. (N.D.). *The Prevalence Of Heart Disease And Stroke: Assessing The Fatality And Remedial Strategies For Adults And Elderly People In Akwa Ibom State*.

Papadopoulos, P., Soflano, M., Chaudy, Y., Adejo, W., & Connolly, T. M. (2022). A Systematic Review Of Technologies And Standards Used In The Development Of Rule-Based Clinical Decision Support Systems. *Health And Technology*, *12*(4), 713–727.

Parato, A. G., Parato, V. M., Fedacko, J., & Magomedova, A. (2024). Global Burden Of Causes Of Death And Life Expectancy With Reference To Cardiovascular Diseases. *World Heart Journal*, *16*(2), 103–115.

Rane, N., Choudhary, S., & Rane, J. (2024). Ensemble Deep Learning And Machine Learning: Applications, Opportunities, Challenges, And Future Directions. *Opportunities, Challenges, And Future Directions (May 31, 2024)*.

Rasheed, K., Qayyum, A., Ghaly, M., Al-Fuqaha, A., Razi, A., & Qadir, J. (2022). Explainable, Trustworthy, And Ethical Machine Learning For Healthcare: A Survey. *Computers In Biology And Medicine*, *149*, 106043.

Siddique, S., Haque, M. A., George, R., Gupta, K. D., Gupta, D., & Faruk, M. J. H. (2023). Survey On Machine Learning Biases And Mitigation Techniques. *Digital*, *4*(1), 1–68.

Subbaswamy, A., Adams, R., & Saria, S. (2021). Evaluating Model Robustness And Stability To Dataset Shift. *International Conference On Artificial Intelligence And Statistics*, 2611–2619.

Tasci, E., Zhuge, Y., Camphausen, K., & Krauze, A. V. (2022). Bias And Class Imbalance In Oncologic Data—Towards Inclusive And Transferrable AI In Large Scale Oncology Data Sets. *Cancers*, *14*(12), 2897.

Tayefi, M., Ngo, P., Chomutare, T., Dalianis, H., Salvi, E., Budrionis, A., & Godtliebsen, F. (2021). Challenges And Opportunities Beyond Structured Data In Analysis Of Electronic Health Records. *Wiley Interdisciplinary Reviews: Computational Statistics*, *13*(6), E1549.

Zhou, C., Dai, P., Hou, A., Zhang, Z., Liu, L., Li, A., & Wang, F. (2024). A Comprehensive Review Of Deep Learning-Based Models For Heart Disease Prediction. *Artificial Intelligence Review*, *57*(10), 263.