Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 3 July 2024, pp: 1034-1042
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Comparative Analysis of Machine Learning Models for Chronic Disease Indicator Classification Using U.S. Chronic Disease Indicators Dataset

Gregorius Airlangga

Information System Study Program, Atma Jaya Catholic University of Indonesia, Indonesia

E-Mail: gregorius.airlangga@atmajaya.ac.id

Received May 14th 2024; Revised Jun 18th 2024; Accepted Jun 23th 2024
Corresponding Author: Gregorius Airlangga

Abstract

The prevalence of chronic diseases poses significant challenges to public health systems worldwide. This study evaluates the performance of four machine learning models—Gradient Boosting Classifier, Support Vector Machine (SVM), Logistic Regression, and Random Forest—in classifying chronic disease indicators using the U.S. Chronic Disease Indicators (CDI) dataset. The models were assessed based on accuracy, precision, recall, F1 score, classification report, and confusion matrix to determine their effectiveness. The Gradient Boosting Classifier outperformed other models with an accuracy of 64.36%, precision of 63.72%, recall of 64.36%, and F1 score of 63.88%. While SVM and Random Forest demonstrated moderate performance, Logistic Regression served as a baseline for comparison. The study highlights the Gradient Boosting Classifier's superiority in handling the complexities of the CDI dataset, suggesting its potential for improving chronic disease prediction and management. Future research should focus on refining these models, addressing class imbalances, and incorporating domain knowledge to enhance interpretability and applicability in real-world scenarios.

Keyword: Chronic Disease Classification, Chronic Disease Indicators, Gradient Boosting, Machine Learning, Support Vector Machine.

1. INTRODUCTION

Chronic diseases, also known as non-communicable diseases (NCDs), have emerged as a leading cause of morbidity and mortality globally, placing immense pressure on healthcare systems and economies [1]–[3]. In the United States, chronic diseases such as heart disease, cancer, and diabetes are responsible for approximately 70% of all deaths, underscoring the urgent need for effective prevention, management, and treatment strategies [4]–[6]. These diseases not only diminish the quality of life for individuals but also contribute significantly to healthcare costs and lost productivity [7]. The growing prevalence of chronic diseases has spurred extensive research efforts aimed at understanding their etiology, risk factors, and optimal intervention approaches [8]. In this context, the application of machine learning (ML) techniques offers promising avenues for enhancing the accuracy and efficiency of chronic disease prediction and management [9]. The integration of machine learning into healthcare analytics has the potential to revolutionize the early detection and diagnosis of chronic diseases [10]. By leveraging vast amounts of health-related data, ML algorithms can identify patterns and correlations that may elude traditional statistical methods [11]. This capability is particularly valuable given the multifactorial nature of chronic diseases, which often involve complex interactions between genetic, environmental, and lifestyle factors [12]. The literature on machine learning in healthcare is extensive, encompassing a wide range of approaches and methodologies. Studies have demonstrated the efficacy of various ML models, including decision trees, support vector machines, neural networks, and ensemble methods, in predicting the onset and progression of chronic diseases [13].

One of the critical challenges in chronic disease research is the accurate classification of disease indicators from large and heterogeneous datasets [14]. Traditional statistical methods, while valuable, often fall short in handling the complexity and high dimensionality of such data. Machine learning techniques, on the other hand, can effectively manage and analyze large-scale datasets, uncovering subtle patterns and relationships that are crucial for disease prediction and prognosis [15]. The state-of-the-art ML models, such as random forests, gradient boosting machines, and deep learning architectures, have shown remarkable performance in various healthcare applications, from image-based diagnostics to electronic health record (EHR) analysis [16]–[19]. Despite the significant advancements in this field, there are still notable gaps and challenges that need to be addressed. One major gap is the generalizability of ML models across different populations and healthcare settings. Many studies focus on specific cohorts or datasets, limiting the

DOI: https://doi.org/10.57152/malcom.v4i3.1403

applicability of their findings to broader contexts. Additionally, issues related to data quality, such as missing values, imbalanced classes, and noise, pose substantial hurdles to the development of robust ML models [20]. Another critical challenge is the interpretability of complex ML models, which is essential for gaining the trust of healthcare professionals and ensuring the clinical utility of the predictions.

Our research aims to contribute to this evolving field by developing and evaluating machine learning models for the classification of chronic disease indicators using a comprehensive dataset from the U.S. Chronic Disease Indicators (CDI) database. The CDI dataset, maintained by the Centers for Disease Control and Prevention (CDC), provides a rich source of information on chronic disease prevalence, risk factors, and preventive measures across the United States [21]. By applying advanced ML techniques to this dataset, we seek to enhance the accuracy and interpretability of chronic disease predictions, thereby supporting more effective disease management strategies. The primary goal of our research is to build and compare the performance of different ML models, including Random Forest, Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting Classifier, in predicting chronic disease indicators. We will systematically evaluate these models based on various performance metrics, such as accuracy, precision, recall, and F1 score, to identify the most effective approach for this task. Additionally, we aim to address common data quality issues, such as missing values and imbalanced classes, through appropriate preprocessing and resampling techniques. Our study will also explore the interpretability of the models by examining feature importance and model explanations.

A critical aspect of our research is the gap analysis, which highlights the limitations of existing studies and identifies areas for improvement. One significant gap is the lack of comprehensive evaluations of different ML models on the CDI dataset, which encompasses a wide range of chronic disease indicators and risk factors. Furthermore, many existing studies do not adequately address the challenges posed by data quality issues, leading to potential biases and inaccuracies in their predictions. Our research seeks to fill these gaps by providing a thorough and systematic evaluation of multiple ML models on the CDI dataset, incorporating robust data preprocessing and validation techniques. The contribution of our research lies in its comprehensive approach to chronic disease classification using machine learning. By systematically comparing different ML models and addressing key data quality issues, we aim to provide valuable insights into the most effective techniques for predicting chronic disease indicators. Our findings will have practical implications for healthcare practitioners and policymakers, supporting more accurate and timely interventions for chronic disease prevention and management. Additionally, our study will contribute to the broader field of healthcare analytics by advancing the understanding of ML applications in chronic disease research.

The remaining structure of this journal article is organized as follows: Section 2 details the research materials and methodology, including data preprocessing, model selection, and evaluation procedures. Section 3 discusses the results of our experiments, comparing the performance of different ML models and analyzing their implications. Finally, Section 4 concludes the article by summarizing the key findings, discussing the limitations of our study, and suggesting directions for future research.

2. MATERIALS AND METHOD

The research method section delineates the systematic procedures employed in our study to develop and evaluate machine learning models for the classification of chronic disease indicators as presented in the Figure 1. This section encompasses data collection, preprocessing, feature selection, model selection, model training and evaluation, hyperparameter tuning, and performance metrics. Our approach ensures a rigorous and reproducible methodology aimed at achieving robust and interpretable results.

2.1. Data Collection

The dataset used in this study is the U.S. Chronic Disease Indicators (CDI) dataset, which is publicly available and maintained by the Centers for Disease Control and Prevention (CDC). The CDI dataset comprises a comprehensive collection of data on chronic disease prevalence, risk factors, and preventive measures across the United States. It includes various attributes such as demographic information, health behaviors, and clinical measures, providing a rich and detailed source of information suitable for machine learning applications in chronic disease research. The CDI dataset is structured to include several key types of information. Demographic information includes attributes such as age, sex, race, and geographic location, which help in understanding the distribution of chronic diseases across different population groups. Health behaviors encompass information on lifestyle factors such as smoking status, physical activity, and dietary habits, which are critical risk factors for many chronic diseases. Clinical measures include data such as blood pressure, cholesterol levels, and body mass index (BMI), which provide direct indicators of health status and chronic disease risk, the dataset also can be downloaded from [22].

The dataset is organized in a tabular format, where each row represents an observation (e.g., an individual or a population group) and each column represents a specific attribute. For instance, the dataset may include columns for *Age*, *Sex*, *Race*, *SmokingStatus*, *PhysicalActivityLevel*, *BloodPressure*, *Cholesterol*, and various other measures relevant to chronic disease monitoring. Mathematically, the dataset can be represented as a matrix (X) of dimensions $(n \times m)$, where (n) is the number of observations and (m) is the number of attributes. Each element (x_{ij}) in the matrix corresponds to the value of the (j)-th attribute for the (i)-th observation. The target variable (y), indicating the type of data value, can be represented as a vector of length (n) as presented in the equation 1.

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{1}$$

The CDI dataset's comprehensive coverage and detailed information make it particularly valuable for machine learning applications in chronic disease research. The diversity of attributes allows for the exploration of complex interactions between demographic factors, health behaviors, and clinical measures. For example, understanding how lifestyle factors such as physical activity and diet interact with clinical measures like blood pressure and cholesterol can provide deeper insights into the risk factors and prevention strategies for chronic diseases.

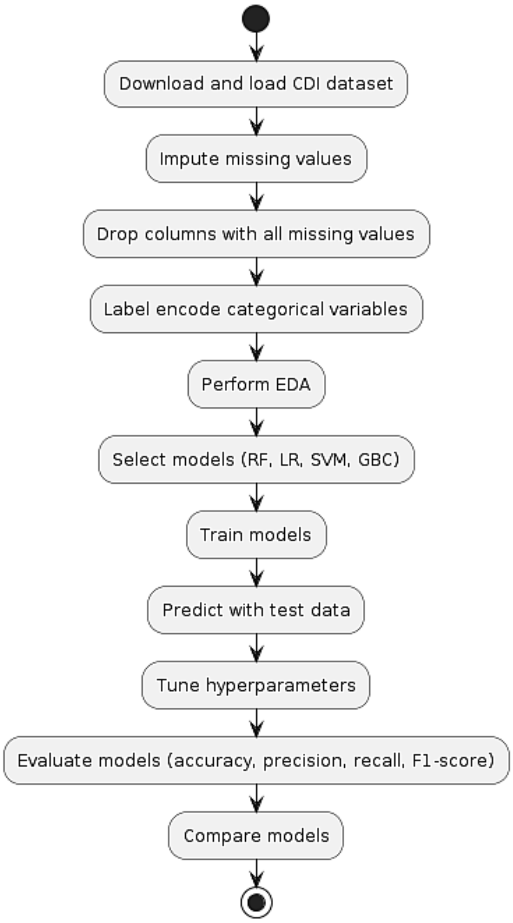
Download and load CDI dataset
Impute missing values
Drop columns with all missing values
Label encode categorical variables
Perform EDA
Select models (RF, LR, SVM, GBC)
Train models
Predict with test data
Tune hyperparameters
Evaluate models (accuracy, precision, recall, F1-score)
Compare models

Figure 1. Research Methodology

2.2. Data Acquisition

The dataset is accessed through the CDC's online data repository, which ensures that it is up-to-date and standardized. The data acquisition process involves downloading the dataset in CSV format from the CDC's website, loading the dataset into a pandas DataFrame for preprocessing and analysis, and performing basic exploratory data analysis (EDA) to understand the structure and characteristics of the dataset, including summary statistics and data visualization. Furthermore, a summary of the CDI dataset provides an overview of the key attributes and their distributions. Summary statistics such as mean, median, standard deviation, and range are calculated for numerical attributes as presented in the equation (2) – (5), while frequency distributions are determined for categorical attributes. This initial exploration helps in identifying potential data quality issues such as missing values and outliers, which are addressed in the data preprocessing stage.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{2}$$

$$\text{Media n}(x) = x_{\left(\frac{n+1}{2}\right)} \tag{3}$$

$$\sigma_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{4}$$

$$\text{Range } (x) = x_{\max} - x_{\min} \tag{5}$$

2.3. Data Preprocessing

Given the nature of the CDI dataset, some attributes contain missing values. To address this, we employed the following strategies, for numerical attributes such as *DataValue*, *LowConfidenceLimit*, and *HighConfidenceLimit*, missing values were imputed using the median of the respective columns. This approach is robust to outliers and provides a central tendency measure for imputation. The imputation can be expressed mathematically as follows: Let (x_i) represent the (i)-th value of a numerical attribute. If (x_i) is missing, it is replaced by the median $(\tilde{x})$ of the attribute, calculated as presented in the equation 6.

$$\tilde{x} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right) & \text{if } n \text{ is even} \end{cases} \tag{6}$$

Where (n) is the number of non-missing values in the attribute. Columns with all missing values were dropped from the dataset to avoid introducing noise or irrelevant information into the models. This step ensures that the dataset only contains relevant and complete information for the analysis. In addition, the dataset contains several categorical variables that need to be converted into a numerical format suitable for machine learning algorithms. We used label encoding for this purpose where each categorical variable was transformed into an integer representation using the *LabelEncoder* from the *sklearn* library. Let *C* be a categorical variable with unique categories $(\{c_1, c_2, \ldots, c_k\})$. The *LabelEncoder* maps each category (c_j) to an integer *(j)* where $C \rightarrow \{1, 2, \ldots, k\}$. Furthermore, a dictionary of label encoders was maintained for each categorical column to ensure consistent encoding across different splits of the dataset. Formally, for a categorical variable (C), let (L_C) be the label encoder such that $L_C(c_j) = j$ for $j \in \{1,2, \ldots, k\}$, this encoding ensures that the categorical variables are consistently transformed into a numerical format, making them suitable for machine learning algorithms while preserving the categorical nature of the data. Consider a small example where we have a dataset with the following attributes: *Age*, *Sex*, *Race*, *DataValue*, *LowConfidenceLimit*, and *HighConfidenceLimit*. Suppose the *DataValue* and *LowConfidenceLimit* columns have missing values. The preprocessing steps would be as follows: First, we impute the missing values in the *DataValue* and *LowConfidenceLimit* columns with their respective medians as presented in the equation 7 and 8.

$$\text{DataValue}_{\text{missing}} \rightarrow \text{Data}\widetilde{\text{aV}}\text{alue}, \tag{7}$$

$$\text{LowConfidenceLimit}_{\text{missing}} \rightarrow \text{LowCon}\widetilde{\text{fide}}\text{nceLimit} \tag{8}$$

Next, we drop any columns that are entirely missing. In this example, assume none of the columns are entirely missing. Then, we apply label encoding to the categorical variables *Sex* and *Race*. Let *Sex* have categories $\{Male, Female\}$ and *Race* have categories $\{White, Black, Asian\}$. The label encoders map these categories to integers Sex → {1 (Male), 2 (Female)}, Race → {1 (White), 2 (Black), 3 (Asian)}. By following these preprocessing steps, we ensure that the dataset is clean, complete, and ready for the subsequent stages of machine learning model training and evaluation.

2.4. Model Selection

We selected a diverse set of machine learning models to compare their performance in classifying chronic disease indicators. The chosen models include the Random Forest Classifier, Logistic Regression, Support Vector Machine (SVM), and Gradient Boosting Classifier. Each model offers unique advantages and methodologies, allowing us to evaluate their effectiveness in this specific context. The Random Forest Classifier is an ensemble method that constructs multiple decision trees and combines their outputs to improve accuracy and reduce overfitting. Each decision tree is trained on a bootstrap sample of the dataset, and the final prediction is obtained by aggregating the predictions of all individual trees, typically through majority voting for classification tasks. The key concept of the Random Forest is to reduce the variance of the model without significantly increasing the bias. Mathematically, the prediction $(\hat{y})$ for an input (x) is given by $\hat{y} = \text{majority\_vote}\big(T_1(x), T_2(x), \ldots, T_B(x)\big)$ where (T_i) represents the $(\ i\ )$-th decision tree in the forest, and $(\ B\ )$ is the total number of trees. Second model that we used is logistic regression, the logistic Regression is a linear model used for binary classification that can be extended to multiclass problems using techniques such as one-vs-rest (OvR) or multinomial logistic regression.

The model estimates the probability that a given input (x) belongs to a particular class by applying the logistic function to a linear combination of the input features. For binary classification, the probability $(P(\ y = 1 \mid x\ ))$ is modeled as: $P(\ y = 1 \mid x\ ) = \sigma(w^{\top x} + b) = \frac{1}{1+e^{-(w^{\top x}+b)}}$, where (w) is the weight vector, $(\ b\ )$ is the bias term, and $(\ \sigma)$ is the logistic sigmoid function. The model parameters (w) and $(\ b\ )$ are learned by maximizing the likelihood function or equivalently minimizing the cross-entropy loss. Third model that we used is Support Vector Machine (SVM), SVM is a robust classifier that finds the optimal hyperplane for separating classes in a high-dimensional space. The objective of SVM is to maximize the margin between the closest points of the different classes, known as support vectors. For a linearly separable case, the decision function is defined as $f(x) = w^{\top x} + b$, the optimization problem for SVM can be formulated as presented in the equation 9.

$$\min_{\boldsymbol{w},b} \quad \frac{1}{2}|\boldsymbol{w}|^2 \quad \text{subject to} \quad y_i(\boldsymbol{w}^{\top x_i} + b) \geq 1, \forall i \tag{9}$$

where (y_i) are the class labels. For non-linearly separable data, kernel functions are used to transform the input space into a higher-dimensional space where a linear separator can be found. Last model that we used is Gradient Boosting Classifier, the method is an ensemble technique that builds sequential models, each correcting the errors of its predecessor, to achieve high predictive performance. The model is built by iteratively adding weak learners, typically decision trees, to the ensemble. Each new tree is trained to approximate the negative gradient of the loss function with respect to the model's predictions. Mathematically, the model prediction $(\hat{y_i}) for\ aninput(x_i)\ at\ iteration(m)$ is updated as $\widehat{y_i^{(m)}} = \widehat{y_i^{(m-1)}} + \nu h_m(x_i)$, where $(\widehat{y_i^{(m-1)}})$ is the prediction from the previous iteration, $(h_m(x_i))$ is the new weak learner added at iteration $(\ m\ )$, and $(\ \nu)$ is the learning rate. The weak learner (h_m) is trained to minimize the residual errors of the current ensemble.

2.5. Model Training and Evaluation

The selected models were trained on the training dataset and evaluated on the testing dataset through a series of methodical steps to ensure robust performance assessment. The process involved training, prediction, and evaluation metrics, which are detailed below. The models were trained using the *fit* method on the training data. Let X_{train} and (y_{train}) denote the training feature matrix and the training target vector, respectively. Each model *(M)* is trained by minimizing an objective function $(\mathcal{L})$, which could be specific to the algorithm used. The training process can be represented as $M = fit(X_{train}, y_{train})$, for instance, in the case of logistic regression, the objective is to minimize the negative log-likelihood (or equivalently, the cross-entropy loss) to estimate the model parameters (w) and $(\ b\ )$. Furthermore, the trained models were then used to predict the target variable on the testing data. Let (X_{ts}) represent the testing feature matrix. The predictions $(\hat{y})$ are obtained as $\hat{y} = M(X_{\text{ts}})$. This step involves applying the learned model parameters to the input features of the testing data to produce the predicted labels.

2.6. Evaluation Metrics

The model performance was evaluated using multiple metrics to provide a comprehensive assessment of the model's predictive capabilities. These metrics include accuracy, precision, recall, F1 score, classification report, and confusion matrix. The accuracy metric measures the proportion of correctly predicted instances out of the total instances. It is given by $\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$ where $(\ TP\ )$ is the number of true positives, $(\ TN\ )$ is the number of true negatives, $(\ FP\ )$ is the number of false positives, and $(\ FN\ )$ is the number of false negatives. Secondly, we used precision where the precision measures the proportion of true positive predictions out of all positive predictions. It is defined as $\text{Precision} = \frac{TP}{TP+FP}$. Thirdly, we used recall, where recall, also known as sensitivity, measures the proportion of true positive predictions out of all actual positives. It is given by $\text{Recall} = \frac{TP}{TP+FN}$, Lastly we used F1 score where the F1 score is the harmonic mean of precision and recall, providing a single metric that balances both. It is calculated as $\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$.

3. RESULTS AND DISCUSSION

In this section, we present the results of our machine learning models applied to the U.S. Chronic Disease Indicators (CDI) dataset. The models evaluated include Gradient Boosting Classifier, Support Vector Machine (SVM), Logistic Regression, and Random Forest. We compare their performance using metrics such as accuracy, precision, recall, F1 score, classification report, and confusion matrix. These metrics provide a comprehensive assessment of each model's effectiveness in classifying chronic disease indicators.

3.1. Gradient Boosting Classifier Performance

The Gradient Boosting Classifier achieved an accuracy of 0.6436, with a precision of 0.6372, recall of 0.6436, and an F1 score of 0.6388. The classification report indicates varied performance across different classes. For instance, class 0 (presumably a specific chronic disease indicator) had a precision of 0.49, recall of 0.52, and F1 score of 0.51, highlighting moderate predictive capability. Notably, the model performed exceptionally well in classifying classes 4, 7, 9, 10, and 11, with precision and recall values often exceeding 0.70 as presented in the table 2. This suggests that the model is particularly effective for certain chronic disease indicators but may require improvement for others. The Gradient Boosting Classifier's superior performance, as indicated by its highest accuracy of 0.6436, demonstrates its ability to handle the complexities of the dataset. This model's strength lies in its iterative boosting process, which sequentially reduces errors from previous iterations, thereby enhancing prediction accuracy. The high precision and recall for classes 4, 7, 9, 10, and 11 suggest that the model is particularly adept at distinguishing these chronic disease indicators, possibly due to their distinctive feature patterns. The variability in performance across different classes highlights the need for further refinement. For instance, the lower performance in classes 0, 1, and 2 suggests that these classes may have overlapping feature distributions or that the model requires additional tuning, such as adjusting learning rates or increasing the number of estimators.

3.2. Support Vector Machine (SVM) Performance

The SVM model achieved an accuracy of 0.4797, with a precision of 0.5240, recall of 0.4797, and an F1 score of 0.4750. The classification report shows significant variability in performance across different classes. For instance, class 0 had a precision of 0.09 and a recall of 0.54, indicating poor precision but relatively high recall. This disparity suggests that the SVM model may be generating many false positives for certain classes, thus affecting overall precision. Conversely, the model exhibited strong performance for classes 3, 6, 9, 10, and 11, with recall values often close to or at 1.00, indicating high sensitivity for these classes as presented in the table 3. The SVM model's performance is characterized by a moderate overall accuracy of 0.4797. Its high recall values for classes 3, 6, 9, 10, and 11 indicate that the model is highly sensitive to these classes, effectively identifying most true positive cases. However, the low precision for class 0 (0.09) and high recall (0.54) imply that the model generates many false positives, thus affecting its precision. This issue can be addressed by optimizing the regularization parameter (C) and kernel type to better balance precision and recall. The SVM's ability to perform well in certain classes suggests that the chosen kernel function effectively maps the input features into a higher-dimensional space where these classes are more separable. However, the model's overall performance indicates that it may not be the best choice for this particular dataset, given the substantial variability in class performance.

3.3. Logistic Regression Performance

The Logistic Regression model attained an accuracy of 0.5244, with a precision of 0.5248, recall of 0.5244, and an F1 score of 0.5178. The classification report reveals significant inconsistencies, particularly in classes 3, 10, and 11, where the model failed to make correct predictions, resulting in precision and recall values of 0.00. However, the model showed relatively balanced performance for classes 1, 4, 6, 7, and 9, with precision and recall values around 0.40 to 0.65 as presented in the table 4. These results suggest that while Logistic Regression can provide reasonable performance for some classes, it struggles significantly with others. Logistic Regression's performance, with an accuracy of 0.5244, reflects its capability to provide a baseline comparison for other models. The model's balanced performance in certain classes, such as 1, 4, 6, 7, and 9, indicates that it can handle linearly separable classes reasonably well. However, the poor performance in classes 3, 10, and 11, where precision and recall are zero, suggests that these classes may not be linearly separable, leading to misclassification. Improving Logistic Regression's performance could involve using more advanced techniques such as regularization adjustments or employing polynomial features to capture non-linear relationships in the data. However, given its relatively lower performance compared to Gradient Boosting, it may not be the optimal model for this dataset.

3.4. Random Forest Classifier

The Random Forest Classifier achieved an accuracy of 0.4798, with a precision of 0.4809, recall of 0.4798, and an F1 score of 0.4803. The classification report indicates a mixed performance, with the model showing reasonable effectiveness for classes 1, 2, 4, 6, 7, 8, and 9, where precision and recall values were around 0.40 to 0.72 as presented in the table 5. However, the model struggled with class 0, which had a precision and recall of 0.17, indicating a high number of false positives and false negatives. The model performed exceptionally well for classes 10 and 11, achieving perfect scores in both precision and recall. The Random Forest Classifier's performance, with an accuracy of 0.4798, reflects its ensemble nature, which helps in reducing overfitting and capturing complex interactions in the data. The model's strength is evident in its reasonable performance across most classes, particularly classes 1, 2, 4, 6, 7, 8, and 9. The high performance in classes 10 and 11, where precision and recall are perfect, indicates that the model can handle certain classes with distinctive features very well. However, the poor performance in class 0, with precision and recall both at 0.17, indicates that the model struggles with classes having overlapping feature distributions. Further tuning, such as adjusting the number of trees, maximum depth, and minimum samples per split, could enhance the model's performance.

3.5. Discussion

The results indicate that the Gradient Boosting Classifier outperforms the other models in terms of accuracy, precision, recall, and F1 score as presented in the table 1. This model's iterative boosting process effectively reduces errors and enhances prediction accuracy, making it the most suitable choice for classifying chronic disease indicators in the CDI dataset. However, the variability in performance across different classes suggests that further refinement is necessary to improve the model's robustness. The SVM and Random Forest models show moderate performance, with notable strengths in specific classes. Their performance can be improved through hyperparameter tuning and by employing advanced techniques such as ensemble methods for SVM or feature importance analysis for Random Forest to better understand and enhance model predictions.

Logistic Regression provides a reasonable baseline but struggles with non-linearly separable classes. Its performance could be improved with more advanced techniques, but it remains less effective compared to Gradient Boosting. In conclusion, the Gradient Boosting Classifier's superior performance makes it the preferred model for this study. Future work should focus on further tuning and refining this model to enhance its performance across all classes. Additionally, exploring hybrid models that combine the strengths of different algorithms could provide further improvements in classifying chronic disease indicators. One limitation of this study is the inherent class imbalance in the dataset, which can affect model performance. Techniques such as oversampling, undersampling, or synthetic data generation (e.g., SMOTE) could be employed to address this issue. Additionally, the models' interpretability could be enhanced by using techniques such as SHAP (SHapley Additive exPlanations) values to understand feature contributions to model predictions.

Table 1. Model Performance Summary

Metric	Gradient Boosting	SVM	Logistic Regression	Random Forest
Accuracy	0.6436	0.4797	0.5244	0.4798
Precision	0.6372	0.524	0.5248	0.4809
Recall	0.6436	0.4797	0.5244	0.4798
F1 Score	0.6388	0.475	0.5178	0.4803

Table 2. Classification Report - Gradient Boosting

Class	Precision	Recall	F1-Score	Support
0.0	0.49	0.52	0.51	1305.0
1.0	0.53	0.46	0.49	15196.0
2.0	0.56	0.44	0.49	4787.0
3.0	0.33	0.27	0.3	11.0
4.0	0.73	0.79	0.76	1079.0
5.0	0.62	0.59	0.6	1721.0
6.0	0.79	0.73	0.76	1171.0
7.0	0.71	0.76	0.73	25850.0
8.0	0.54	0.59	0.56	5585.0
9.0	0.77	0.83	0.8	5576.0
10.0	1.0	1.0	1.0	36.0
11.0	1.0	1.0	1.0	32.0

Table 3. Classification Report - SVM

Class	Precision	Recall	F1-Score	Support
0.0	0.09	0.54	0.15	1305.0
1.0	0.46	0.23	0.3	15196.0
2.0	0.43	0.24	0.31	4787.0
3.0	0.62	0.91	0.74	11.0
4.0	0.59	0.43	0.5	1079.0

Class	Precision	Recall	F1-Score	Support
5.0	0.56	0.51	0.54	1721.0
6.0	0.49	0.94	0.64	1171.0
7.0	0.61	0.59	0.6	25850.0
8.0	0.5	0.3	0.37	5585.0
9.0	0.45	0.93	0.61	5576.0
10.0	0.65	1.0	0.79	36.0
11.0	1.0	1.0	1.0	32.0

Table 4. Classification Report - Logistic Regression

Class	Precision	Recall	F1-Score	Support
0.0	0.4	0.16	0.23	1305.0
1.0	0.44	0.47	0.46	15196.0
2.0	0.4	0.28	0.33	4787.0
3.0	0.0	0.0	0.0	11.0
4.0	0.41	0.52	0.46	1079.0
5.0	0.46	0.25	0.33	1721.0
6.0	0.54	0.52	0.53	1171.0
7.0	0.55	0.65	0.6	25850.0
8.0	0.45	0.34	0.39	5585.0
9.0	0.89	0.67	0.76	5576.0
10.0	0.0	0.0	0.0	36.0
11.0	0.0	0.0	0.0	32.0

Table 5. Classification Report - Random Forest

Class	Precision	Recall	F1-Score	Support
0.0	0.17	0.17	0.17	1305.0
1.0	0.28	0.28	0.28	15196.0
2.0	0.4	0.41	0.41	4787.0
3.0	0.62	0.91	0.74	11.0
4.0	0.62	0.64	0.63	1079.0
5.0	0.35	0.34	0.35	1721.0
6.0	0.66	0.64	0.65	1171.0
7.0	0.57	0.57	0.57	25850.0
8.0	0.47	0.46	0.47	5585.0
9.0	0.72	0.72	0.72	5576.0
10.0	1.0	1.0	1.0	36.0
11.0	1.0	1.0	1.0	32.0

4. CONCLUSION

This study aimed to evaluate the performance of various machine learning models in classifying chronic disease indicators using the U.S. Chronic Disease Indicators (CDI) dataset. The models assessed included Gradient Boosting Classifier, Support Vector Machine (SVM), Logistic Regression, and Random Forest Classifier. Through comprehensive evaluation using metrics such as accuracy, precision, recall, F1 score, classification report, and confusion matrix, we identified the strengths and limitations of each model. Among the evaluated models, the Gradient Boosting Classifier demonstrated the best overall performance, achieving the highest accuracy of 0.6436 and balanced precision, recall, and F1 scores. This model's iterative boosting process effectively reduced errors, making it particularly adept at distinguishing between various chronic disease indicators. However, its performance varied across different classes, suggesting the need for further refinement and tuning to improve robustness. The SVM and Random Forest models showed moderate performance with notable strengths in specific classes. Their performance can be enhanced through hyperparameter tuning and the application of advanced techniques. Logistic Regression, while providing a reasonable baseline, struggled with non-linearly separable classes and demonstrated lower overall effectiveness compared to Gradient Boosting. This study highlights the potential of machine learning models, particularly ensemble methods like Gradient Boosting, in accurately classifying chronic disease indicators. However, it also underscores the importance of addressing class imbalances and further refining models to enhance their applicability and robustness. Future work should focus on refining the Gradient Boosting Classifier, exploring hybrid models that combine the strengths of different algorithms, and incorporating domain knowledge to improve model interpretability. Additionally, addressing dataset limitations such as class imbalance and expanding the dataset to include more diverse chronic disease indicators could further enhance model performance and generalizability.

ACKNOWLEDGMENTS

We express our deepest gratitude to Atma Jaya Catholic University of Indonesia for the unwavering support and resources provided throughout the duration of this research project. Our sincere thanks also extend to the Information System Study Program, whose rigorous academic environment and collaborative ethos have been instrumental in facilitating our work.

REFERENCES

[1] F. Luna and V. A. Luyckx, "Why have non-communicable diseases been left behind?," *Asian Bioeth. Rev.*, vol. 12, no. 1, pp. 5–25, 2020.

[2] H. Singh and J. Bharti, "Non-Communicable Diseases and Their Risk Factors," *EAS J Parasitol Infect Dis*, vol. 3, no. 6, pp. 83–86, 2021.

[3] B. Gyawali, P. Khanal, S. R. Mishra, E. van Teijlingen, and D. Wolf Meyrowitsch, "Building strong primary health care to tackle the growing burden of non-communicable diseases in Nepal," *Glob. Health Action*, vol. 13, no. 1, p. 1788262, 2020.

[4] A. Budreviciute, S. Damiati, D. K. Sabir, and R. Kodzius, "Management and prevention strategies for non-communicable diseases (NCDs) and their risk factors," *Front. public Heal.*, vol. 8, p. 574111, 2020.

[5] A. Francis *et al.*, "Chronic kidney disease and the global public health agenda: an international consensus," *Nat. Rev. Nephrol.*, pp. 1–13, 2024.

[6] M. A. Faghy *et al.*, "Cardiovascular disease prevention and management in the COVID-19 era and beyond: an international perspective," *Prog. Cardiovasc. Dis.*, vol. 76, pp. 102–111, 2023.

[7] N. P. F. Pequeno, N. L. de A. Cabral, D. M. Marchioni, S. C. V. C. Lima, and C. de O. Lyra, "Quality of life assessment instruments for adults: a systematic review of population-based studies," *Health Qual. Life Outcomes*, vol. 18, pp. 1–13, 2020.

[8] V. Falanga *et al.*, "Chronic wounds," *Nat. Rev. Dis. Prim.*, vol. 8, no. 1, p. 50, 2022.

[9] Z. Ahmed, K. Mohamed, S. Zeeshan, and X. Dong, "Artificial intelligence with multi-functional machine learning platform development for better healthcare and precision medicine," *Database*, vol. 2020, p. baaa010, 2020.

[10] M. Sarker, "Revolutionizing healthcare: the role of machine learning in the health sector," *J. Artif. Intell. Gen. Sci. ISSN 3006-4023*, vol. 2, no. 1, pp. 36–61, 2024.

[11] A. S. Morrow, A. D. Campos Vega, X. Zhao, and M. M. Liriano, "Leveraging machine learning to identify predictors of receiving psychosocial treatment for Attention Deficit/Hyperactivity Disorder," *Adm. Policy Ment. Heal. Ment. Heal. Serv. Res.*, vol. 47, no. 5, pp. 680–692, 2020.

[12] J. V. S. Guerra, M. M. G. Dias, A. J. V. C. Brilhante, M. F. Terra, M. Garcia-Arevalo, and A. C. M. Figueira, "Multifactorial basis and therapeutic strategies in metabolism-related diseases," *Nutrients*, vol. 13, no. 8, p. 2830, 2021.

[13] J. Yang, X. Ju, F. Liu, O. Asan, T. S. Church, and J. O. Smith, "Prediction for the risk of multiple chronic conditions among working population in the United States with machine learning models," *IEEE Open J. Eng. Med. Biol.*, vol. 2, pp. 291–298, 2021.

[14] Z. Nenova and J. Shang, "Chronic disease progression prediction: Leveraging case-based reasoning and big data analytics," *Prod. Oper. Manag.*, vol. 31, no. 1, pp. 259–280, 2022.

[15] F. Nazi and T. Abbas, "Harnessing Machine Learning for Cancer Subtype Classification: Precision Medicine Applications," *J. Environ. Sci. Technol.*, vol. 2, no. 2, pp. 72–82, 2023.

[16] V. A. Lepakshi, "Machine learning and deep learning based AI tools for development of diagnostic tools," in *Computational Approaches for Novel Therapeutic and Diagnostic Designing to Mitigate SARS-CoV-2 Infection*, Elsevier, 2022, pp. 399–420.

[17] D. Painuli, S. Bhardwaj, and others, "Recent advancement in cancer diagnosis using machine learning and deep learning techniques: A comprehensive review," *Comput. Biol. Med.*, vol. 146, p. 105580, 2022.

[18] S. Dixit, A. Kumar, and K. Srinivasan, "A Current Review of Machine Learning and Deep Learning Models in Oral Cancer Diagnosis: Recent Technologies, Open Challenges, and Future Research Directions," *Diagnostics*, vol. 13, no. 7, p. 1353, 2023.

[19] M. Kirola, M. Memoria, A. Dumka, and K. Joshi, "A comprehensive review study on: optimized data mining, machine learning and deep learning techniques for breast cancer prediction in big data context," *Biomed. Pharmacol. J.*, vol. 15, no. 1, pp. 13–25, 2022.

[20] T. Emmanuel, T. Maupong, D. Mpoeleng, T. Semong, B. Mphago, and O. Tabona, "A survey on missing data in machine learning," *J. Big data*, vol. 8, pp. 1–37, 2021.

[21] E. Finn, F. L. Andersson, and M. Madin-Warburton, "Burden of Clostridioides difficile infection (CDI)-a systematic review of the epidemiology of primary and recurrent CDI," *BMC Infect. Dis.*, vol. 21, no. 1, p. 456, 2021.

[22] Jainaru, "Chronic Disease Indicators." 2023.