Supervised Machine Learning Approach for Predicting Cardiovascular Complications Risk in Patients with Diabetes Mellitus

1st Arief Purnama Muharram
Master Program of Informatics
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
23521013@std.stei.itb.ac.id

2nd Fahmi Sajid
Master Program of Informatics
School of Electrical Engineering and Informatics
Institut Teknologi Bandung
Bandung, Indonesia
23522028@std.stei.itb.ac.id

2023 International Conference on Electrical Engineering and Informatics (ICEEI) | 979-8-3503-8129-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICEEI59426.2023.10346215

***Abstract*—Diabetes mellitus, particularly type-2 diabetes, remains a prevalent health issue, raising concerns due to its associated risk of complications. Among these, cardiovascular complications pose a significant threat, exhibiting high morbidity and mortality rates. Health screening plays a pivotal role in stratifying the risk levels of diabetes patients, facilitating proactive measures to prevent the progression of complications. As such, the primary objective of this study is to develop a predictive model system for assessing cardiovascular risk in diabetes patients. Our study used the Cardiovascular Disease dataset and conducts experiments with various supervised machine learning algorithms, such as Naive Bayes, decision tree, random forest, AdaBoost, and XGBoost. The results reveal that ensemble learning algorithms based on boosting, particularly AdaBoost and XGBoost, outperform other supervised machine learning methods. However, even with the best performance achieved using the dataset, the accuracy stands at 0.71, and the F-1 score is 0.69, which is still acceptable for screening purposes. Although these results provide valuable insights, indicating individuals at higher risk for cardiovascular complications in diabetes, further improvements are needed to enhance early prevention strategies.**

***Index Terms*—diabetes mellitus, cardiovascular complications, machine learning, Naive Bayes, decision tree, random forest, AdaBoost, XGBoost**

I. Introduction

Diabetes mellitus (or diabetes) is a group of metabolic disorders characterized by chronic high blood sugar levels [1], [2]. Type-2 diabetes mellitus (T2DM) is the most common type of diabetes, with more than 90% of people with diabetes having T2DM [2]. T2DM is closely related to an unhealthy lifestyle, with lack of physical activity, high sugar intake, and sedentary behavior among the causes that increase the risk of T2DM [1]. One of the concerns with diabetes is the potential complications that may arise later in life. Among the complications that can arise from diabetes, vascular complications (or complications happened in blood vessels) require special attention.

Vascular complications in diabetes can be divided into microvascular and macrovascular complications [1], [3]. Microvascular complications include retinopathy, nephropathy, and neuropathy. Macrovascular complications include coronary heart disease, cardiomyopathy, arrhythmia, stroke, and peripheral artery disease. These vascular complications arise from chronic inflammation resulting from high blood sugar levels, leading to damage in the blood vessels. [3]. The risk is even greater and occurs more quickly in people who do not manage their condition well [4]. One vascular complication that draws our attention is cardiovascular complications.

Cardiovascular complications are a term for a group of diseases or disorders that occur in the heart and blood vessels, such as coronary heart disease, stroke, peripheral arterial disease, and heart failure [1]. The risk of cardiovascular complications in patients with diabetes can be as high as 32.2% [5]. Furthermore, patients with diabetes experience a 2 to 8-fold increase in cardiovascular mortality [4], making cardiovascular complications the leading cause of morbidity and mortality in people with diabetes [5].

Due to the high morbidity and mortality rates associated with the diabetes complications, early screening approaches are essential. Health screening is intended to identify individuals who are at a higher risk, allowing for early prevention actions to be taken. Therefore, there is a need to create a predictive system to make health screening efficient and effective.

This study aimed to develop a predictive model system for cardiovascular complication risk in diabetes patients using a supervised machine learning approach. The goal of this study is to provide a system that can assist in the effort to screen for cardiovascular complication risks in diabetes patients efficiently based on an artificial intelligence approach.

The contribution of this study can be summarized as follows:

- Identifying the optimal supervised machine learning model for predicting the risk of cardiovascular complications in diabetes patients.
- Developing a predictive system capable of assessing the risk of cardiovascular complications in diabetes patients.

II. Related Work

The issue arising from the dangers of cardiovascular complications in diabetes patients has encouraged researchers to develop a machine learning method for predicting the risk of these complications in such patients. Several studies have been conducted to develop supervised machine learning models for predicting cardiovascular complications in diabetes patients.

One of the early studies in this field was conducted by Giardina et al [6]. In their study, they employed genetic algorithms (GA) and Weighted k-Nearest Neighbors (WkNN) to classify the presence of coronary heart disease in patients with T2DM. Furthermore, they compared the Random Initialization (RI), Feature Ranking using Correlation Coefficient (FRCC), and Knowledge-based Initialization (KI) techniques as part of the GA initialization process. The results showed that the GA/WkNN/KI approach yielded the best results compared to other combinations. The study achieved a sensitivity of 44.20% and specificity of 79.48%.

Another study was conducted by Miao et al [7]. In their research, they utilized the Framingham Heart Study dataset along with Support Vector Machine (SVM) and k-Nearest Neighbors (KNN) algorithms to predict cardiovascular disease in T2DM patients. Their study resulted in an accuracy rate of 96.5% and a recall rate of 89.8% when using the SVM algorithm. Additionally, they achieved an accuracy rate of 96.9% and a recall rate of 92.9% when using the KNN algorithm. However, the study did not explicitly specify the inclusion of only diabetes patients in the sample population for developing models aimed at predicting cardiovascular diseases in T2DM diabetes.

III. Methodology

A. Dataset

In this study, we used the Cardiovascular Disease dataset [8]. The Dataset contains attributes related to the history of cardiovascular disease obtained from patients' medical records. The dataset consists of 70,000 data points, which are comprised of 11 attributes as features and 1 attribute as the target. The feature attributes are divided into 3 types: objective, examination, and subjective.

- **Objective features** are factual data (age, $height$, $weight$, and $gender$).
- **Examination features** are medical examination results and are objective in nature (ap_hi (systolic blood pressure), ap_lo (diastolic blood pressure), $cholesterol$, and $glucose$).
- **Subjective features** are information provided by the patient and are subjective in nature ($smoke$ (smoking history), $alco$ (alcohol consumption history), and $active$ (physical activity)).

To obtain patient data with a history of diabetes from the dataset, we used an assumption-based approach. Our assumption was that patients with a "normal" $glucose$ attribute belong to the non-diabetic group, as diabetes is characterized by elevated blood glucose levels. While it is possible for diabetic patients to have normal blood glucose readings, in this particular dataset, we considered those cases as non-diabetic due to the absence of specific attributes differentiating diabetic and non-diabetic patients. Therefore, based on this assumption, we excluded patient data with a "normal" $glucose$ attribute from the dataset.

Furthermore, to ensure data cleanliness and eliminate outliers, we established the following exclusion criteria, which were thoroughly discussed with experts in the field:

- Rule out the $height$ attribute below 120 (cm) and above 200 (cm).
- Rule out the $weight$ attribute below 40 (kg) and above 140 (kg).
- Rule out the ap_hi attribute below 60 (mmHg) and above 210 (mmHg).
- Rule out the ap_lo attribute below 40 (mmHg) and above 140 (mmHg).

B. Feature Engineering

In this study, we performed feature engineering to enhance the dataset. Three specific feature engineering techniques were applied: BMI calculation, BMI status determination, and hypertension classification.

*1) **Body Mass Index (BMI) Calculation**:* We calculate the body mass index of the patient using the following standard formula.

$$\text{BMI } (kg/m^2) = \frac{\text{Body Weight } (kg)}{\text{Body Height } (m)^2} \tag{1}$$

The calculation results were rounded to two decimal places only and stored as bmi attribute.

TABLE I
WHO BMI Classification

Status	**BMI** (kg/m^2)
Underweight	< 18.5
Normal	$18.5 - 24.9$
Overweight	$25.0 - 29.9$
Obese	≥ 30.0

TABLE II
BMI Classification Used in Our Study

Status	**BMI** (kg/m^2)
Underweight	< 18.5
Normal	$18.5 - 24.9$
Overweight	≥ 25.0

*2) **BMI Status Determination**:* Based on the BMI data obtained, we then conducted feature engineering on BMI status. We used the BMI classification by World Health Organization (WHO) [9], which consists of 4 categories: underweight, normal, overweight, and obese (Table I). However, we reduced the classification into 3 categories, namely underweight, normal, and overweight for simplification. We classify obesity in the same category as overweight since it falls under the broader overweight classification (Table II). The results were then stored as bmi_status attribute.

*3) **Hypertension Classification**:* We determined hypertension classification based on the American Heart Association (AHA) hypertension classification (Table III) [10]. The AHA classifies our blood pressure into five categories: normal, elevated, hypertension stage I, hypertension stage II, and hypertension crisis. However, in this study, we reduced the "elevated" category into "normal" because "elevated" is still considered within the normal range. The results were then stored as ht_stage attribute (Table IV).

TABLE III
AHA HYPERTENSION CLASSIFICATION

Classification	Systolic BP (mmHg)	and/or	Diastolic BP (mmHg)
Normal	Less than 120	and	Less than 80
Elevated	120-129	and	Less than 80
Hypertension Stage I	130-139	or	80-89
Hypertension Stage II	140 or higher	or	90 or higher
Hypertension Crisis	Higher than 180	and/or	Higher than 120

TABLE IV
HYPERTENSION CLASSIFICATION USED IN OUR STUDY

Classification	Systolic BP (mmHg)	and/or	Diastolic BP (mmHg)
Normal	Less than 130	and	Less than 80
Hypertension Stage I	130-139	or	80-89
Hypertension Stage II	140 or higher	or	90 or higher
Hypertension Crisis	Higher than 180	and/or	Higher than 120

TABLE V
DESIGN OF EXPERIMENTS

Parameter	Variations	Count
Machine learning algorithms	Naive Bayes, Decision Tree, Random Forest, AdaBoost, and XGBoost	5
Hyperparameter tuning	With and without GridSearchCV	2
	Total of Experiments	**10**

C. Design of Experiments

The objective of this study's experiment is to find the best-performing classification model to predict the risk of cardiovascular disease in diabetic patients. In this experiment, we varied various supervised machine learning algorithms, including Naive Bayes, decision tree, random forest, AdaBoost, and XGBoost. Furthermore, the experiment was conducted both with and without using GridSearchCV to determine the best hyperparameters for the model to be compared. Validation was performed using the hold-out validation scheme with a test dataset ratio of 0.33. The performance measurement matrices used were accuracy and F1-Score. Table V summarizes the experimental strategies we used.

For experiments that did not use GridSearchCV, we used the default parameters of each supervised machine learning algorithm in the library. In developing our model, we used Scikit-Learn library [11] version 0.24.1 and trained it using Google Colab.

D. Deployment

The model will be deployed in the form of a web application. The web application consists of a form with fields that can be filled with user data. We use Streamlit version 1.8.1 for our deployment method.

IV. RESULT

A. Feature Selection

TABLE VI
SUMMARY STATISTICS OF THE DATASET

Attribute	Mean±Std	Min	Max	Corr
age	54.26±6.4	39	64	0.18
bmi	29.08±5.7	13.52	63.98	0.16
ap_hi	130.56±17.8	60	210	0.34
ap_lo	83.08±9.8	40	120	0.28
Attribute	**Category**	**Count**	**Perc**	**Corr**
ht_stage	Normal (1)	1399	0.14	0.3
	HT Stage I (2)	5625	0.55	
	HT Stage II (3)	3188	0.31	
	HT Crisis (4)	26	Very Small	
$cholesterol$	Normal (1)	3645	0.36	0.17
	Above (2)	2723	0.27	
	Well Above (3)	3870	0.38	

After we performed dataset cleaning, we obtained 10,238 data as the dataset to be used. From the dataset, there were 6,169 "yes" labels and 4,069 "no" labels of the history of cardiovascular disease. Based on the provided labels, we considered the "yes" label to indicate a high-risk of cardiovascular disease, while the "no" label indicated a low-risk. This assumption was grounded in empirical evidence. Specifically, we recognized that just because the original label was "have cardiovascular" did not necessarily mean that it must signify the presence of "cardiovascular disease." Conversely, the label "doesn't have cardiovascular disease" did not imply that the individual would never develop cardiovascular disease in the future, but rather indicated a lower risk instead.

We then used the feature correlation method to select the features to be used (Figure 1). Based on the analysis of feature correlations, 6 following attributes were selected with the highest correlation values, where all of these attributes were in line with diabetes complications as risk factors. [1] (Table VI):

- ap_hi (systolic blood pressure) with a correlation value of 0.34.
- ht_stage (hypertension stage) with a correlation value of 0.3.

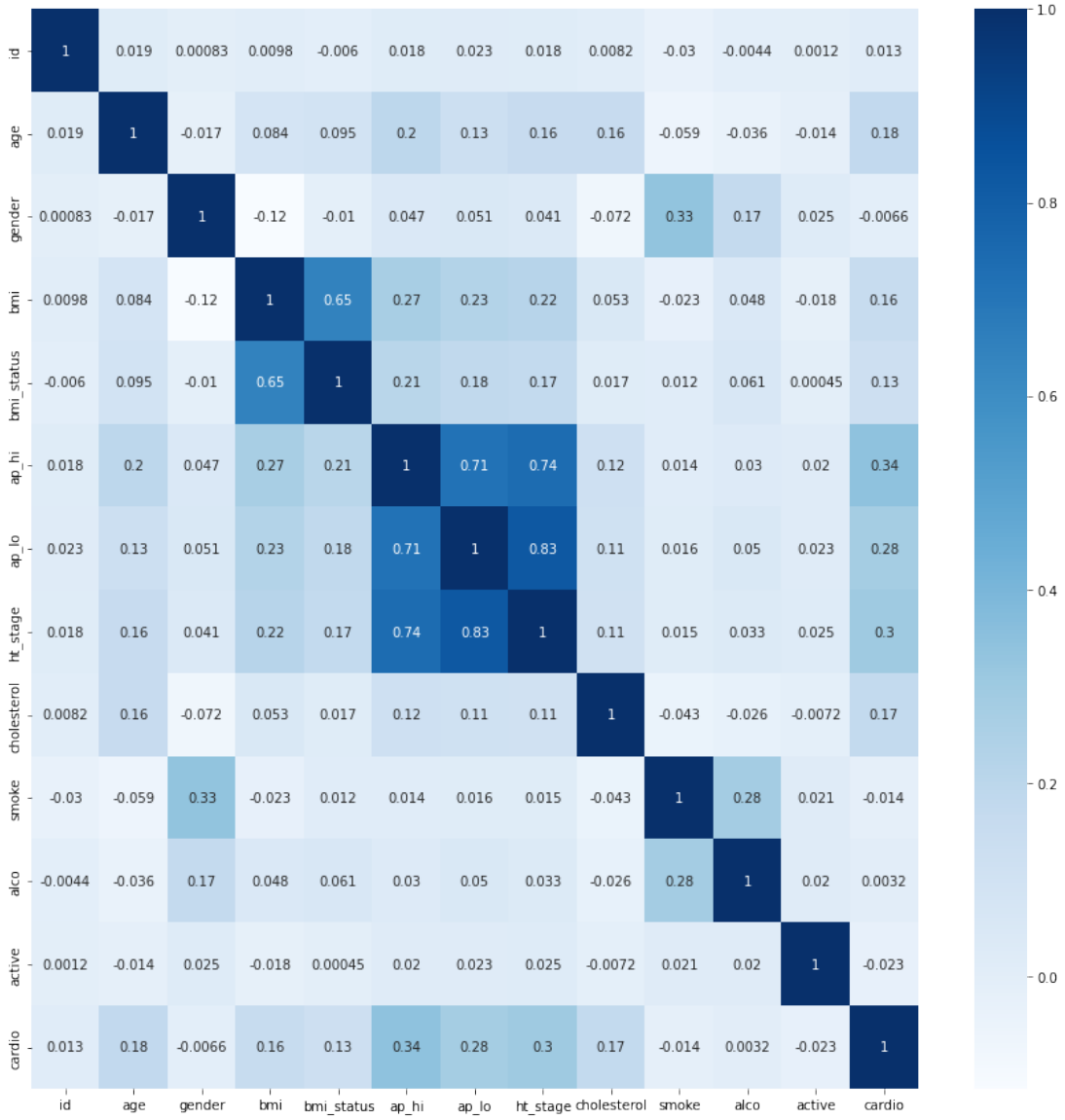

Fig. 1. Feature correlation matrix

- ap_lo (diastolic blood pressure) with a correlation value of 0.28.
- age with a correlation value of 0.18.
- $cholesterol$ with a correlation value of 0.17.
- bmi with a correlation value of 0.16.

Although the number of data points in the HT crisis category was very small, we decided not to merge it with the other HT stages. This decision was based on its different clinical significance and implications.

B. Models Development

In each algorithm, we conducted two types of experimental scenarios: default parameters and GridSearchCV. Firstly, we trained the model using the training dataset, and the results can be seen in Table VII for default parameters and Table VIII for GridSearchCV. Next, we evaluated it using the test dataset, and the results can be seen in Table IX for default parameters and Table X for GridSearchCV.

The experimental results showed that the AdaBoost and XGBoost algorithms performed the best in predicting the risk of cardiovascular complications. The AdaBoost and XGBoost algorithms were able to achieve accuracy performances up to 0.71 and macro average F-1 scores up to 0.69, respectively. The F-1 score analysis with macro average was preferred because the dataset used was imbalanced, with the target label of high-risk class being relatively more prevalent than the low-risk class.

The similarity between the AdaBoost and XGBoost algorithms is that both are members of the ensemble learning method. Ensemble learning is a machine learning approach that uses a combination of more than one machine learning algorithm [12]. Ensemble learning algorithms excel in various

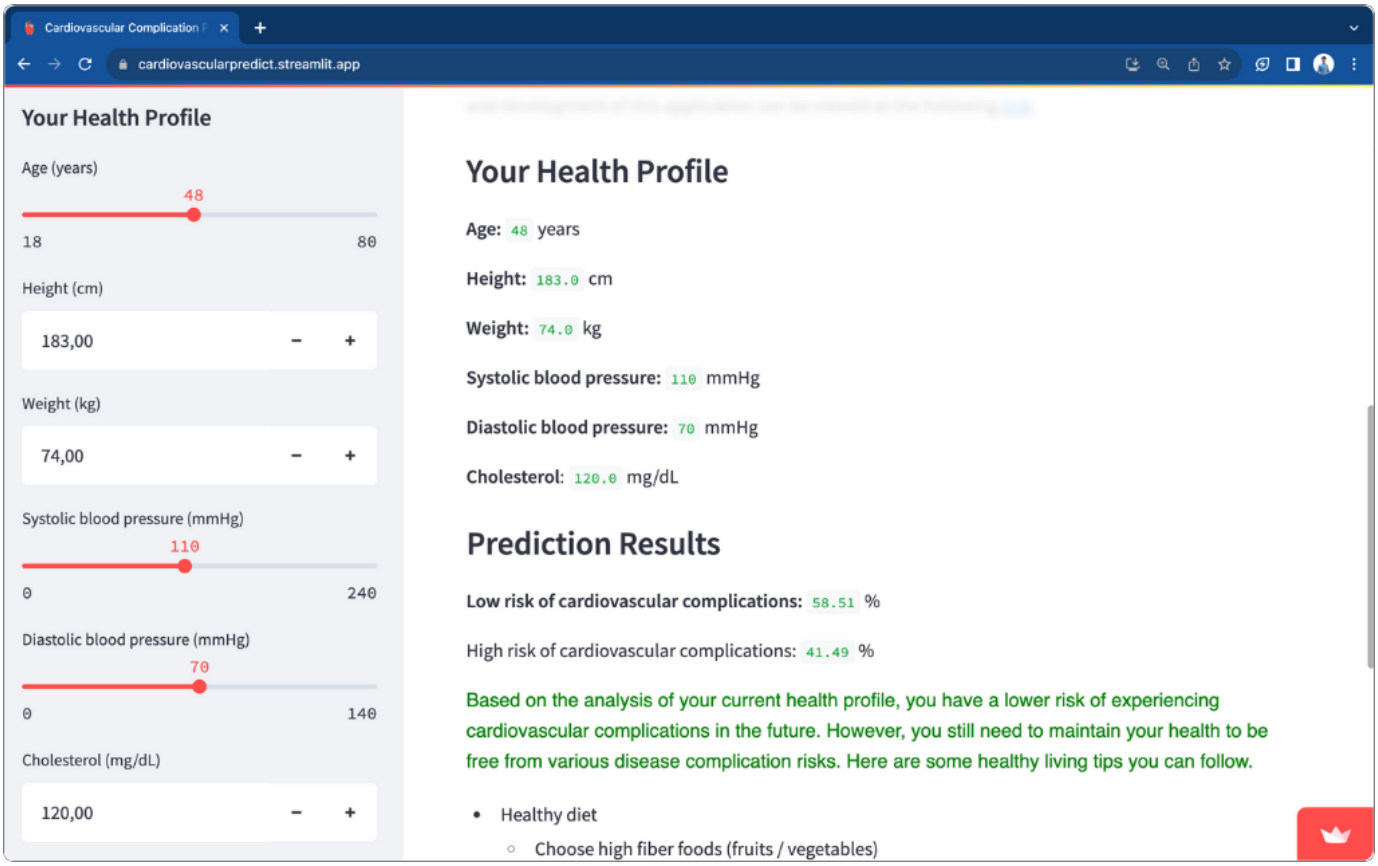

Fig. 2. Our system demonstration

TABLE VII
TRAIN RESULTS - DEFAULT HYPERPARAMETER

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	**69%**	**68%**	**68%**	**68%**
Decision tree	61%	60%	60%	60%
Random forest	64%	63%	62%	64%
AdaBoost	67%	66%	66%	66%
XGBoost	66%	67%	66%	66%

TABLE VIII
TRAIN RESULTS - GRIDSEARCHCV

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	67%	**67%**	65%	65%
Decision tree	**68%**	66%	66%	66%
Random forest	**68%**	**67%**	66%	66%
AdaBoost	**68%**	**67%**	**67%**	**67%**
XGBoost	**68%**	**67%**	66%	66%

TABLE IX
TEST RESULTS - DEFAULT HYPERPARAMETER

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	69%	68%	68%	68%
Decision tree	60%	58%	58%	58%
Random forest	66%	64%	64%	64%
AdaBoost	**71%**	69%	**69%**	**69%**
XGBoost	**71%**	**70%**	**69%**	**69%**

TABLE X
TEST RESULTS - GRIDSEARCHCV

Model	Accuracy	Precision	Recall	F1-score
Naive Bayes	70%	68%	67%	67%
Decision tree	70%	68%	68%	68%
Random forest	70%	68%	68%	68%
AdaBoost	**71%**	69%	**66%**	**69%**
XGBoost	**71%**	**70%**	68%	**69%**

tasks because they use the performance of multiple algorithms, which can reduce the error rate that may occur.

While the ensemble algorithms used in this study (XGBoost and AdaBoost) outperformed the Naive Bayes algorithm, it's important to note that their performance differences were not significant. Furthermore, the best-performing model in this study fell significantly short in terms of accuracy when compared to the study conducted by Miao et al [7]. Specifically, our study achieved an accuracy rate of 71%, whereas Miao et al. reported a much higher accuracy of 96.9%.

This subpar model performance can be attributed to several potential factors, including:

- **Dataset Selection:** Our study's dataset, particularly the way we assumed the population of patients with T2DM within it, may have influenced the results.
- **Feature Engineering and Imbalanced Data Handling:**

The manner in which we handled feature engineering, imbalanced datasets, and re-annotation might have impacted the model's performance.

- **Algorithm Choice:** The choice of machine learning algorithms used in this study may not have been the most suitable for this particular problem.

It's important to note that in our study, we did not address the issue of imbalanced datasets; we used the dataset as-is without any modification. Additionally, when comparing our results to those of Miao et al. [7], it's essential to further explore whether they used the expected population of T2DM patients as the baseline for their study.

The use of the GridSearch method for hyperparameter tuning in this experiment did not significantly affect the results. In the simple tree algorithm family, namely decision tree and random forest, GridSearch was able to improve the F-1 score by a few points (decision tree: 0.58 to 0.68; random forest: 0.64 to 0.68). However, in the more advanced ensemble models based on boosting, namely AdaBoost and XGBoost, the use of GridSearch did not improve the F-1 score (AdaBoost: 0.69; XGBoost: 0.69). On the other hand, in the Naive Bayes algorithm, the use of GridSearch tended to decrease the F-1 score (Naive Bayes: 0.68 to 0.67). One of the main factors that could be considered for the low F-1 score generated, despite hyperparameter tuning, is the quality of the dataset used.

C. Model Deployment

We deployed our model using the Streamlit library. To use it, users only need to enter their data, and the results will appear on the screen (Figure 2). We used soft classification to allow users to see the confidence level of the prediction results. The source code for our system is accessible through https://github.com/fahmisajid/CardiovascularPredict.

V. Conclusion

The ensemble learning algorithms based on boosting, namely AdaBoost and XGBoost, are able to provide better prediction performance in terms of accuracy and F-1 score compared to other supervised machine learning algorithms such as Naive Bayes, decision tree, and random forest. However, the best performance that can be achieved using the Cardiovascular Disease Dataset is an accuracy of 0.71 and an F-1 score of 0.69. This can be caused by several factors such as: (1) the dataset selection; (2) the feature engineering and imbalance dataset handling; and (3) the algorithm used. Suggestions for further research that can be recommended are the use of deep learning algorithms and the careful selection of the high-quality data.

Acknowledgement

We would like to acknowledge and give our thanks to the lecturer of IF5171 Machine Learning for DSAI course, Master Program of Informatics, School of Electrical Engineering, Institut Teknologi Bandung, who has given us the knowledge and basic understanding of machine learning. This task was originally a part of the authors' coursework, and we are thankful for the opportunity and challenges to apply the knowledge gained in class to a real project.

References

[1] Pengurus Besar Perkumpulan Endokrinologi Indonesia, *Konsensus Pengelolaan dan Pencegahan Diabetes Melitus Tipe 2 di Indonesia 2019*. Jakarta, Indonesia: Pengurus Besar Perkumpulan Endokrinologi Indonesia, 2019.

[2] International Diabetes Federation, *IDF Diabetes Atlas*. Brussels, Belgia: International Diabetes Federation, 2022. [Online]. Available: https://www.diabetesatlas.org. Accessed: October 8, 2022.

[3] M. J. Fowler, "Microvascular and Macrovascular Complications of Diabetes," in *Clinical Diabetes*, vol. 26, no. 2, pp. 77-82, Apr. 2008. doi: 10.2337/diaclin.26.2.77.

[4] P. Farmaki, C. Damaskos, N. Garmpis, A. Garmpi, S. Savvanis, and E. Diamantis, "Complications of Type 2 Diabetes Mellitus," in *Current Cardiology Reviews*, vol. 16, no. 4, pp. 249-251, 2020. doi: 10.2174/1573403X1604201229115531.

[5] T. R. Einarson, A. Acs, C. Ludwig, and U. H. Panton, "Prevalence of cardiovascular disease in type 2 diabetes: a systematic literature review of scientific evidence from across the world in 2007-2017," in *Cardiovascular Diabetology*, vol. 17, no. 83, 2018. doi: 10.1186/s12933-018-0728-6

[6] M. Giardina, F. Azuaje, P. McCullagh and R. Harper, "A Supervised Learning Approach to Predicting Coronary Heart Disease Complications in Type 2 Diabetes Mellitus Patients," *Sixth IEEE Symposium on BioInformatics and BioEngineering (BIBE'06)*, Arlington, VA, USA, 2006, pp. 325-331, doi: 10.1109/BIBE.2006.253297.

[7] L. Miao, X. Guo, H. T. Abbas, K. A. Qaraqe and Q. H. Abbasi, "Using Machine Learning to Predict the Future Development of Disease," *2020 International Conference on UK-China Emerging Technologies (UCET)*, Glasgow, UK, 2020, pp. 1-4, doi: 10.1109/UCET51115.2020.9205373.

[8] *Cardiovascular Disease Dataset*, Kaggle, 2021. [Online]. Available: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset. Accessed: September 13, 2022.

[9] C. B. Weir and A. Jan, "BMI Classification Percentile and Cut Off Points," in *StatPearls*. StatPearls. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK541070/. Accessed: September 18, 2023.

[10] American Heart Association, "Understanding Blood Pressure Readings." Heart.org. Available: https://www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings. Accessed: May 12, 2023.

[11] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," in *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.

[12] G. Brown, "Ensemble Learning," in *Sammut, C., Webb, G.I. (eds) Encyclopedia of Machine Learning*. Springer, Boston, MA, 2011, pp. 312–320, doi: 10.1007/978-0-387-30164-8_252.