JURNAL EMACS
[Engineering, MAthematics and Computer Science] Vol.7 No.1 January 2025: 1-6
e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i1.12846

Leveraging Support Vector Machines and Ensemble Learning for Early Diabetes Risk Assessment: A Comparative Study

Hafizh Ash Shiddiqi[1*], Karli Eka Setiawan[2], Renaldy Fredyan[3]
[1,2]Computer Science Department, School of Computer Science,
Bina Nusantara University, Jakarta, Indonesia 11480
[3]Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei 106335, Taiwan
hafizh.shiddiqi@binus.ac.id; karli.setiawan@binus.ac.id; D11315803@mail.ntust.edu.tw

*Correspondence : hafizh.shiddiqi@binus.ac.id

Abstract – *Currently, diabetes is a hidden, serious threat to human lifestyles through daily food and drink, which has become a formidable global health challenge. As a contribution, this study suggests a way to use machine learning to find people with diabetes by looking at certain health parameters. It does this by using different Support Vector Machine (SVM)-based models, such as different SVMs with different kernels, such as linear, polynomial, radial basis function, and sigmoid kernels; different ensemble bagging with SVM; and different ensemble stacking with various SVM models. The findings demonstrated that utilizing a single SVM model with a linear kernel, ensemble bagging with a linear SVM, and ensemble stacking with different SVM models yielded the most accurate results, achieving 95% accuracy in both diabetes presence and absence. This lends credence to the idea that the incorporation of a linear kernel has the potential to improve the accuracy of determining whether or not diabetic illness is present.*

Keywords: Diabetes, Prediction, Support Vector Machine, Kernels, Ensemble Learning

I. INTRODUCTION

As a chronic non-communicable disease, diabetes has become a formidable global health challenge. The rising prevalence of diabetes emphasizes the critical need for extensive research initiatives to address the significant threats it poses to human health, impacting approximately 425 million individuals worldwide by 2020 (Intelligence and Neuroscience, 2023).

Diabetes prevalence has surged to concerning levels, particularly among aging populations. In 2019, it was reported that 19.3% of individuals aged 65 to 99 were diagnosed with diabetes, highlighting the considerable burden this condition imposes on older adults (Sinclair et al., 2020). These statistics underscore the critical need for intensified research efforts to mitigate the adverse impact of diabetes on public health.

Undiagnosed diabetes poses serious risks, including a greater risk of cardiac stroke, diabetic nephropathy, and various other severe complications. Consequently, early identification is critical for effectively managing diabetes and alleviating its negative health consequences (Kaur & Kumari, 2022). Chronic elevation of blood sugar levels in diabetes can result in severe complications, including kidney damage, vision impairment, cardiovascular diseases, and even premature mortality (Prasetyo et al., 2024). Therefore, timely identification and intervention are crucial for averting the progression of diabetes-related Complications.

One of the fundamental freedoms that society must recognize is the right to health (Setiawan et al., 2023). This research offered a comparative study of machine learning (ML) models to assist health professionals in performing their tasks and saving individuals from diabetes illness at an early stage, which became our research contribution (Aryawibowo et al., 2023). Using the early-stage diabetes risk prediction dataset, this study examined the

capabilities of different SVM-based models, including the SVM model with various kernels such as linear, polynomial, RBF, and sigmoid, as well as ensemble bagging and ensemble stacking. We picked SVM-based ML models because we were curious. The research article is organized as follows: Related works are discussed in the next section, and the executed research methodology is completely covered in Section 3. Section 4 displays and explains the experiment findings, followed by the conclusion and future work.

II. METHODS

This study was conducted by applying ensemble methods to the Support Vector Machine (SVM) machine learning model. The SVM models involved in this study consist of several types of kernels, including Linear Kernel, Polynomial Kernel, Radial Basis Function (RBF) Kernel, and Sigmoid Kernel. The ensemble approaches employed include Ensemble Learning with Bagging and Stacking. In conclusion, this study will identify accuracy, precision, recall, and F1-score values as evaluation metrics.

2.1. Dataset

The dataset used in this study was obtained from the UCI Machine Learning Repository. This dataset contains sign and symptom data of newly diagnosed or at-risk diabetic patients, collected through direct questionnaires from patients at Sylhet Diabetes Hospital in Sylhet, Bangladesh, with medical approval. The dataset comprises 519 instances with 16 parameters that can be utilized to predict diabetes presence. The parameters are as follows: age, gender, polyuria, polydipsia, sudden weight loss, weakness, polyphagia, genital thrush, visual blurring, itching, irritability, delayed healing, partial paresis, muscle stiffness, alopecia, and obesity.

2.2. Support Vector Machine (SVM)

Support Vector Machines (SVM) are primarily known as a classification approach, but they can also be employed for regression problems. SVM effectively handles multiple continuous and categorical variables. The core idea of SVM is to identify a maximum margin hyperplane (MMH) that optimally divides the dataset into distinct classes. To achieve this, SVM constructs a hyperplane in multidimensional space and generates the optimal hyperplane in an iterative manner aimed at minimizing classification error (Cortes & Vapnik, 1995).

2.2.1. Linear Kernel

A linear kernel can be used as normal dot product any two given observations. The product between two vectors is the sum of the multiplication of each pair of input values.

2.2.2. Polynomial Kernel

A polynomial kernel is a more generalized form of the linear kernel. The polynomial kernel can distinguish curved or nonlinear input space.

2.2.3. Radial Basis Function (RBF) Kernel

The Radial basis function kernel is a popular kernel function commonly used in support vector machine classification. RBF can map an input space in infinite dimensional space.

2.2.4. Sigmoid Kernel

Sigmoid kernel in SVM particularly for transforming input data into higher-dimensional space to make it possible to perform linear separation in that space.

2.3. Ensemble Learning

Ensemble learning is an approach that builds learning algorithms from training data obtained through a set of learning algorithms that have been combined to solve the same problem (Zhang & Ma, 2012).

2.3.1. Ensemble Learning with Bagging

Bagging is a common ensemble method for training individual learners on at random parts of the training dataset (Ngo et al., 2022). Ensemble learning with bagging might be referred to as self-sufficient base learner training (González et al., 2020).

As shown in Figure 1 - 4, the inner of ensemble learning with bagging comprised of many comparable models that functioned independently without influence from other weak learners, however in this study, SVM was used as the weak learner. This model used bootstrapping, which is data resampling with replacement, so that during image bagging with

e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i1.12846

SVM in Figure 1 - 4, all SVM models trained separate subsets of the dataset. Each SVM model inside the Bagging models is undoubtedly distinct from one another. Because of the Scikit Learn package, the bagging model's outputs are derived from the soft voting outcomes of each weak learner.

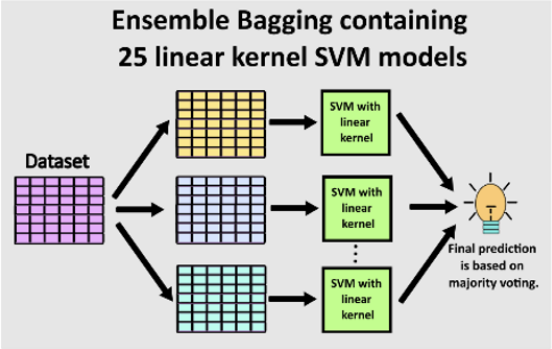

Figure 1. Ensemble Bagging on Linear Kernel SVM

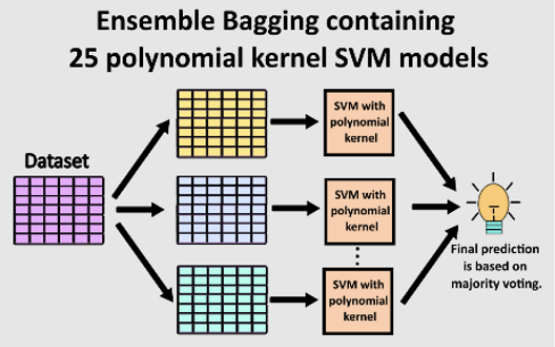

Figure 2. Ensemble Bagging on Polynomial Kernel SVM

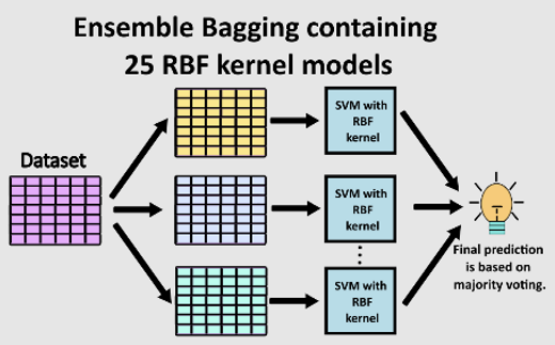

Figure 3. Ensemble Bagging on RBF Kernel SVM

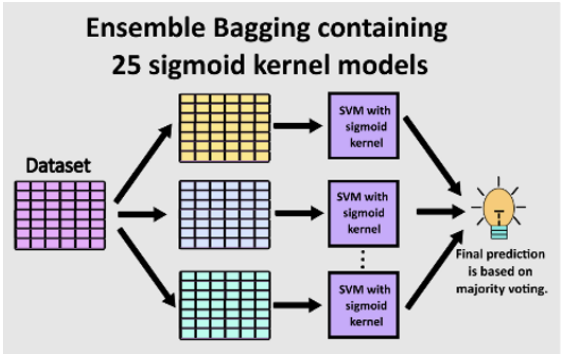

Figure 4. Ensemble Bagging on Sigmoid Kernel SVM

2.3.2. Ensemble Learning with Stacking

Stacking is an ensemble approach, often known as super learning (Kwon et al., 2019). It may produce new data based on projected outcomes using a variety of models as base learners, such random forest, k-nearest neighbors, and support vector machine. This new data is then fed into another predictive model, known as a meta-learner, to determine the final prediction.

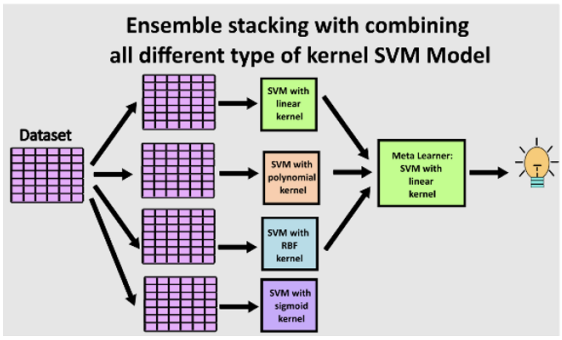

Figure 5. Ensemble Stacking on SVM Kernels

The base learner used in this research was many types of SVM models with different kernel such as linear kernel, polynomial kernel, RBF, and sigmoid.

III. RESULTS AND DISCUSSION

This study performed an investigation about the capability of various support vector machine models with nine different machine learning models. Our first four models were single support vector machine models with different kernels like linear, polynomial, RBF,

and sigmoid as our first, second, third, and fourth models, respectively. Due to our curiosity about implementing many weak learners with 25 SVM models, this research also tried to implement ensemble learning with bagging into our previous first four models as our fifth, sixth, seventh, and eight models. The last model this research proposed was the combination of all SVM models with various kernels as base learners and SVM as meta-learner models into a single ensemble architecture with stacking.

Overall, the fifth, sixth, seventh, eighth, and ninth models used in this research are illustrated in Figure 2. All our models were built using the SCIKIT Learn library in Python programming.

Table 1. Testing result using various SVM-based machine learning models.

Model	Support Vector Machine (SVM)-based Machine Learning Model	Overall		Class 0 (Absence of early stage of diabetes risk)				Class 1 (The Presence of early stage of diabetes risk)			
		Accuracy	Support	Precision	Recall	F1-score	Support	Precision	Recall	F1-score	Support
1	SVM with linear kernel	**0.95**	104	**0.95**	**0.93**	**0.94**	40	**0.95**	**0.97**	**0.96**	64
2	SVM with polynomial kernel	0.62	104	0	0	0	40	0.62	1.00	0.76	64
3	SVM with RBF kernel	0.62	104	0	0	0	40	0.62	1.00	0.76	64
4	SVM with sigmoid kernel	0.55	104	0.41	0.40	0.41	40	0.63	0.52	0.64	64
5	Ensemble Bagging containing 25 linear kernel SVM models	**0.95**	104	**0.95**	**0.93**	**0.94**	40	**0.95**	**0.97**	**0.96**	64
6	Ensemble Bagging containing 25 polynomial kernel SVM models	0.62	104	0	0	0	40	0.62	1.00	0.76	64
7	Ensemble Bagging containing 25 RBF kernel SVM models	0.62	104	0	0	0	40	0.62	1.00	0.76	64
8	Ensemble Bagging containing 25 sigmoid kernel SVM models	0.54	104	0.39	0.35	0.37	40	0.62	0.66	0.64	64
9	Ensemble stacking with combining all different type of kernel SVM Model	**0.95**	104	**0.95**	**0.93**	**0.94**	40	**0.95**	**0.97**	**0.96**	64

e-ISSN: 2686-2573
DOI: 10.21512/emacsjournal.v7i1.12846

This research split our dataset into proportions of 80 and 20 for the training and testing sets, respectively. Based on our experiment result noted in Table 1, in the comparison between four different kernels, the best kernel was the SVM model using a linear kernel with 97 percent recall for the presence of an early stage of diabetes disease risk. Meanwhile, the SVM model with polynomial and RBF kernel failed to detect the health person with 0 percent precision and recall. Both SVM with polynomial and RBF were also bad at predicting the presence of early stages of diabetes disease risk, with 62 percent precision and 100 percent recall. This research also concluded that the implementation of ensemble learning with bagging and stacking seemed to be useless for our data without any improvement, even worse for the SVM model with a sigmoid kernel. We also assumed that ensemble learning using stacking would exactly result in the same result as a single SVM model with a linear kernel because of the capability of the SVM model with a linear kernel used as a base learner and a meta learner. This research concluded that the SVM model with a linear kernel should be enough to solve our binary classification for predicting the presence of early stages of diabetes disease based on our dataset.

IV. CONCLUSION

Eventually, based on our work in investigating the capability of the SVM model in solving binary classification cases for predicting the presence of early stages of diabetes disease, the SVM model with a linear kernel should be enough to be implemented. Because it became our best model by achieving 97 percent recall and 95 percent precision for predicting the presence of early stages of diabetes disease and 93 percent recall and 95 percent precision for predicting the absence of early stages of diabetes disease. Based on our research, we hope that the implementation of machine learning may help with health professional duties and save people from diabetes disease at an early stage. For future research, anyone can expand this research by using other parametric machine learning models such as logistic regression, Naïve Bayes, and others, or other non-parametric machine learning models such as K-Nearest Neighbors, support vector machines, and others.

REFERENCES

Aryawibowo, P., Hidayanto, A. F., Toemali, Y. M., Anderies, Setiawan, K. E., & Gunawan, A. A. S. (2023). Intelligent Monitoring and Diagnosing Capability in Healthcare: Systematic Literature Review. *2023 International Conference on Information Management and Technology (ICIMTech)*, 627–632. https://doi.org/10.1109/ICIMTech59029.2023.10277846

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

González, S., García, S., Del Ser, J., Rokach, L., & Herrera, F. (2020). A practical tutorial on bagging and boosting based ensembles for machine learning: Algorithms, software tools, performance study, practical perspectives and opportunities. *Information Fusion*, *64*, 205–237. https://doi.org/10.1016/j.inffus.2020.07.007

Intelligence and Neuroscience, C. (2023). Retracted: Analysis of Diabetes Clinical Data Based on Recurrent Neural Networks. *Computational Intelligence and Neuroscience*, *2023*(1). https://doi.org/10.1155/2023/9761378

Kaur, H., & Kumari, V. (2022). Predictive modelling and analytics for diabetes using a machine learning approach. *Applied Computing and Informatics*,

18(1/2), 90–100. https://doi.org/10.1016/j.aci.2018.12.004

Kwon, H., Park, J., & Lee, Y. (2019). Stacking ensemble technique for classifying breast cancer. *Healthcare Informatics Research*, *25*(4), 283–288. https://doi.org/10.4258/hir.2019.25.4.283

Ngo, G., Beard, R., & Chandra, R. (2022). Evolutionary bagging for ensemble learning. *Neurocomputing*, *510*, 1–14. https://doi.org/10.1016/j.neucom.2022.08.055

Prasetyo, S. Y., Setiawan, K. E., & Shiddiqi, H. A. (2024). Assessing the Efficacy of Artificial Neural Networks for Diabetes Risk Prediction. *2024 2nd International Symposium on Information Technology and Digital Innovation (ISITDI)*, 108–112. https://doi.org/10.1109/ISITDI62380.2024.10796239

Setiawan, K. E., Kurniawan, A., Chowanda, A., & Suhartono, D. (2023). Clustering models for hospitals in Jakarta using fuzzy c-means and k-means. *Procedia Computer Science*, *216*, 356–363. https://doi.org/10.1016/j.procs.2022.12.146

Sinclair, A., Saeedi, P., Kaundal, A., Karuranga, S., Malanda, B., & Williams, R. (2020). Diabetes and global ageing among 65–99-year-old adults: Findings from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Research and Clinical Practice*, *162*, 108078. https://doi.org/10.1016/j.diabres.2020.108078

Zhang, C., & Ma, Y. (2012). Ensemble Machine Learning. In *Ensemble Machine Learning*. Springer US. https://doi.org/10.1007/978-1-4419-9326-7