Machine Learning Ensemble Techniques for Prediction of Diabetes

Toh Kar Ming[1], Samuel-Soma M. Ajibade[2], Muhammed Basheer Jasser[3], Anwar P.P. Abdul Majeed[4], Kehinde Ehineni[5], Verena Atud Achick-Muyu[6]

[1,2,3,4]Department of Data Science and Artificial Intelligence, School of Engineering and Technology, Sunway University, No. 5, Jalan Universiti, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia

[2]Research Centre for Nanomaterials and Energy Technology (RCNMET), Sunway University, No. 5, Jalan Universiti, Bandar Sunway, 47500 Selangor Darul Ehsan, Malaysia

[5]Industrial Business Unit, Littelfuse, The Galleria, 15 Innovation Blvd, #140, Saskatoon, SK S7N 2X8, Canada

[6]Wolfson Institute of Population Health, Queen Mary University of London, London, UK

[1]21091137@imail.sunway.edu.my, [2]samuelma@sunway.edu.my, [3]basheerj@sunway.edu.my, [4]anwarm@sunway.edu.my
[5]kehineni@littelfuse.com, [6]Verenaachick@yahoo.com,

Abstract*— This study aims to explore hybrid machine learning techniques for the predictive purpose of diabetes disease. The primary objective is to analyze existing supervised machine learning algorithms like K-Nearest Neighbours (KNN), Naïve Bayes (NB), Decision Tree (DT) and so on, that are used for predictive purposes by carrying out extensive research. Through this study, several hybrid machine learning prediction classifiers using supervised techniques and ensemble methods will be studied and evaluated based on their performance matrix of accuracy, precision, recall and F1-score as well. This is to discover which predictive technique is the most suitable for predicting diabetes as well as to evaluate by how much ensemble methods can improve diabetes prediction accuracy. This study utilizes feature selection with Recursive Feature Elimination with Cross Validation (RFECV) to determine the most important features of the dataset, as well as K-Fold Cross Validation to evaluate the performance of all the machine learning techniques used by ensuring that different portions of the dataset is used for training and testing. The results of this study reveal that boosting, by Gradient Boosting (GB), is the best performing ensemble technique at 85.904% accuracy, 70.921% precision, 56.840% recall and 58.439% F1-score. For bagging ensemble techniques, Random Forest (RF) had the highest performance at 85.903% accuracy, 71.019% precision, 56.486% recall and 57.939% F1-score. On the other hand, for stacking ensemble techniques, Decision Tree (DT) + Support Vector Machine (SVM) proved to be the highest performing combination with an accuracy of 85.870%, precision of 70.585%, recall of 57.524% and F1-score of 59.358%. For the base supervised techniques itself, K-Nearest Neighbours (KNN) performed almost as well as GB, with the same accuracy as GB (85.904%), and got a higher precision of 71.050% but had a recall and F1-score that's marginally lower at 56.380% and 57.789% respectively. These findings are significant because, when compared to other existing studies that used the same dataset (Behavioral Risk Factor Surveillance System – 2015 Survey), the performance accuracy of the hybrid machine learning techniques in this study was higher, providing new insights on possible suitable processes taken in this study that can produce better results with the dataset used. To conclude, based on these outcomes we can say that hybrid machine learning techniques generally perform better when compared to supervised machine learning techniques and have the potential to contribute well towards future implementations in the diagnostic medicine field, paving new possibilities to revolutionize traditional diagnostic practices in order to help with the early intervention of diabetes by predictive methods.

Keywords—Machine learning, Ensemble Techniques, prediction, Diabetes, Global health

I. INTRODUCTION

Complications of diabetes can develop over time. The longer a patient goes without regulating their blood sugar, the higher the risk of complications a patient will face which may end up being disabling or even life-threatening. Early detection of diabetes is crucial as it allows for timely intervention of the disease, reducing the complications that are caused by prolonged neglect of the disease such as neuropathy, kidney damage, cardiovascular diseases, and retinopathy [1], as seen in a study done by [2], where it was observed that major benefits were likely to occur from the early diagnosis and treatment of diabetes to reduce the complications of cardiovascular morbidity and mortality. With today's advances in technology, especially in the healthcare sector, machine learning has been frequently discussed as a way to revolutionize how healthcare is carried out. For example, lately machine learning has become more popular than traditional biostatical methods for assessing and integrating vast volumes of complex healthcare data, whether it be for diagnostic purposes or even for the detection of abnormalities in cell tissue images as well [3].

Diabetes is a significant global health issue affecting many worldwide. This condition can lead to several complications such as cardiovascular disease, kidney failure and even neuropathy if not managed properly [4]. Because of these adverse complications and technological advances, machine learning has emerged as a promising tool in the healthcare sector, offering advanced methods for analysing large data sets to predict complications and even disease onset [5]. This works as by leveraging machine learning techniques, it is possible to identify patterns and risk factors associated with diabetes, allowing for timely intervention which will reduce complications and improve quality of life for patients [6]. By examining diabetic patient vitals and lifestyle characteristics, this research attempts to investigate the use of machine learning approaches in diabetes prediction in order to aid with prediction of diabetes in the public to reduce the probability of severe complications for these people later on. Several machine learning techniques used in diagnostic medicine today will be explored to determine which models are most suitable to predict diabetes for the chosen dataset. This exploration aims to enhance the accuracy of the diagnostic software in predicting diabetes, ensuring it delivers reliable and precise results. The remaining of the sections of this research are as follows: Section II discusses the literature review, while methodology is discussed in section III, section IV discusses the result and discussion and section V is the conclusion part of this research paper.

II. LITERATURE REVIEW

The National Institute of Diabetes and Digestive and Kidney Diseases (n.d.) states that diabetes is a common condition brought on by excessively elevated blood glucose levels. Type 1, Type 2 and gestational diabetes are the three most prevalent forms of the disease. Because of an autoimmune reaction or damage to the pancreas that ultimately results in the death of pancreatic bets cells, patients with type 1 diabetes produce little to no insulin which prevents the body from producing the insulin required to control the blood glucose levels. Insulin will always be necessary for these people. However, insulin treatments is usually only necessary for patients with Type 2 diabetes in latter stages. These people have bodily cells that are in some way resistant to insulin. In order to try to get glucose into the body cells, the patients' pancreas must produce more insulin due to the body cell's resistance. Eventually, the pancreas burns out from the high production process, and glucose builds up in the blood stream once more.

A. Supervised Learning Techniques

Classification machine learning consists of algorithms that predict a categorical outcome called classification [7]. The classification algorithms such as decision trees, support vector machines, Naïve Bayes and K-nearest neighbours, are some of the well-known algorithms that are a part of this branch of supervised learning. The algorithms in this category are trained in a process by means of derivation of classification models from training data [8].

i Decision Tree

Decision Trees (DT) are known for dividing data into subsets according to the input feature values [9, 10]. A decision regarding an attribute is represented by each internal node, as shown in fig. 1, and the result is represented by each leaf node. Choosing the optimal attribute to divide the data at each node according to a metric is the first step in building a decision tree. Each path from the root to a leaf in the resulting tree structure reflects a classification rule [11] .

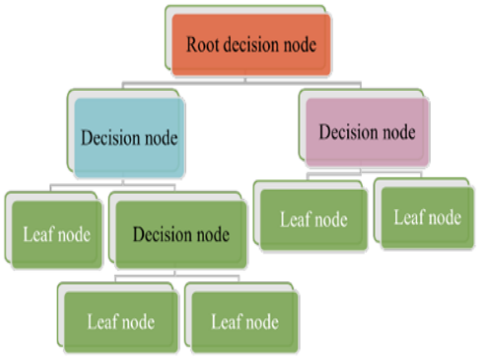

Fig. 1: Decision Tree

ii Support Vector Machine

Supervised learning methods called support vector machines (SVM) examine data for regression and classification [12] . Fig. 2 explains that finding the hyperplane that optimally splits a dataset into classes is how they operate. By optimising the margin between each class's closest data points, or support vectors, the ideal hyperplane is found. SVMs are versatile and effective in high-dimensional spaces because they can handle non-linear classification problems using a variety of kernel functions , including linear, polynomial and radial basis functions [13]. According to [14], each kernel functions has its advantages and disadvantages. For example, the linear kernel function is computationally efficient and performs well in highly dimensional datasets where the data is linearly separable.

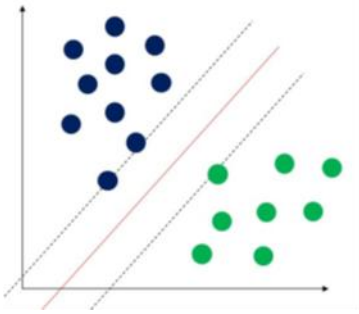

Fig. 2: Support Vector Machine

iii K-Nearest Neighbour

K-Nearest Neighbours (KNN) is a simple, instance-based learning algorithm that classifies a data point based on how its neighbours are classified [15] In KNN, as shown in fig. 3, the 'k' represents the number of nearest neighbours considered to make the prediction [16]. The algorithm assigns the most common class among its k-nearest neighbours to the data point. KNN is useful for diabetes prediction because it can consider multiple factors and similarities in patient data to predict outcomes. It is easy to implement and understand but can be computationally expensive as it requires calculating the distance of each query point to all points in the training dataset [17]. Additionally, it can be sensitive to the choice of 'k' and the distance metric used.

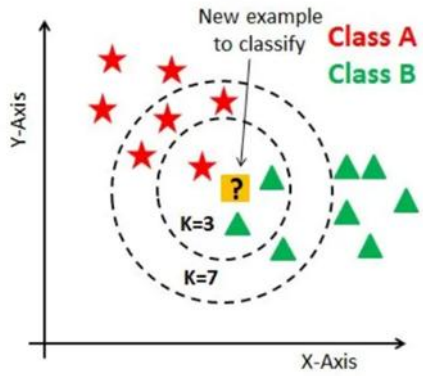

Fig. 3: KNN

B. Unsupervised Learning Techniques

There are three main types of unsupervised learning techniques, clustering, association and dimensionality reduction. Overall, these categories differ primarily in their objectives and methodologies. Clustering focuses on grouping data, association seeks relationships between variables, and dimensionality reduction aims to simplify data without losing essential information [18].

i Clustering

The process of clustering is dividing data into discrete groups or clusters according to the similarities between data points. To make sure that data points in one cluster are more identical to one another than those in other clusters is the goal. Common algorithms include hierarchical clustering, which builds a tree-like structure of nested clusters using either a top-down or bottom-up method, and K-means, which divides data into a predetermined number of groups by minimizing the variation within each cluster. Clustering is widely used in market segmentation, image segmentation and bioinformatics for identifying patterns and grouping similar items [19].

ii Association

Association techniques aim to find interesting relationships, or associations, between variables in large datasets. These techniques are particularly useful in market basket analysis, where the goal is to discover rules that indicate product combinations that frequently co-occur in transactions. This can be seen via the Apriori algorithm which identifies frequent item sets and generates association rules by analysing the co-occurrence of items. Association rule mining helps businesses understand consumer behaviour and design effective marketing strategies [20].

III. METHODOLOGY

A. Data Collection

In phase 1 of the project, a diabetes data set for predictive purposes is gathered from open-sourced repositories such as Kaggle. According to [21], a larger data set was recommended to be used in order to further improve performance. The data set used in the research by [21] had 8 predicator variables, 1 target variable and 768 rows of data, whereas the data set that we'll be using has 15 predicator variables, 1 target variable (they are clarified in Table 1) and 441456 rows of data (Centers for Disease and Control Prevention, n.d.). The data set originally consisted of 330 variables and 441456 rows of data, taken from a 2015 Behavioural Risk Factor Surveillance (BRFSS) survey on Americans collected by the Centers for Disease Control and Prevention (CDC). Fig. 4 explains the phases of our research framework.

Table 1: Features of the Dataset

Column	Feature	Data Type
1	Age Groups	Ordinal
2	Sex	Nominal
3	Race Categories	Nominal
4	BMI Group	Ordinal
5	Level of Education Completed	Ordinal
6	Income Category	Ordinal
7	General Health	Ordinal
8	Physical Activeness	Ordinal
9	Smoking	Ordinal
10	Drinking	Nominal
11	High Blood Pressure	Nominal
12	High Cholesterol	Nominal
13	Stroke	Nominal
14	Coronary Heart Disease or Heart Attack	Nominal
15	Kidney Disease	Nominal
16	Diagnosed Diabetes	Nominal

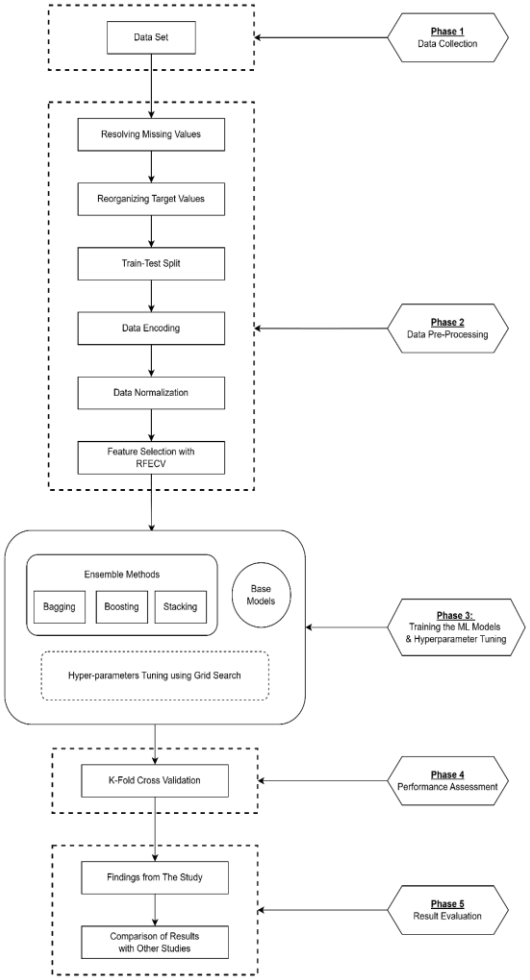

Fig. 4: Research Framework

B. Data Pre-processing

A critical step for machine learning algorithms, data pre-processing, is to be carried out as the quality of input data into the algorithms will have a significant impact on the performance evaluation of used the model in terms of accuracy, precision, recall and fi-score. Thus, to achieve high performance, the preprocessing phase is necessary in machine learning. The data set to be used in this research contains missing values as seen in the form of "not-a-number" (NaN) values. To resolve this, firstly the dataset is checked for "not-a-number" (NaN) values. Any rows containing at least one NaN value will be dropped. Then, each column/variable is checked for values that corresponds with the "Don't know/Not Sure/Refused/Missing" frequency and any rows that contain at least one of that value is dropped. For example, in the "kidney disease" variable, frequency value 9 that represents "Refused" will be dropped as it doesn't help with the prediction of diabetes, and rows that contains these type of value labels in any of its columns/variables will be dropped. As this study will be carrying out prediction of diabetes generally, we will not be considering cases like gestational diabetes and also pre-diabetes and borderline diabetes. This means that the predictions that we

will be doing will generally be Type 1 and Type 2 diabetes which falls under the general "1: Yes" target value category.

Before train-test split is carried out, the target column of the dataset is separated from the feature columns of the dataset in both y and x respectively to ensure that the future models would be able to learn the relationships between inputs (x) and outputs (y). After that, in order to ensure accurate model evaluation, train-test split is carried out with the default setting of 70% of the dataset (178321 rows) being allocated to the training set and 30% of the dataset (76424 rows) being allocated to the testing set.

The purpose of data encoding in the data preprocessing phase is to translate the data of all features to a format that can be effectively utilized by algorithms. When dealing with categorical data its crucial to recognise whether these values represent ordinal or nominal data. If the values are ordinal, label encoding can be used for this case then as it preserves the inherent order of categories [22]. Data normalization is necessary to ensure that no specific feature skews the learning process and affects the model's performance due to its larger range of data, especially in algorithms that rely on distance calculations such as KNN and SVM [23].

C. Model Training and Hyperparameter Tuning

With the completion of data preprocessing, machine learning techniques that can be used for predicting diabetes is then be implemented. The techniques were chosen based on findings from previous diabetes prediction studies with the highest accuracy among the models in Table 2, which are Support Vector Machine (SVM), Naïve Bayes (NB), K-Nearest Neighbour (KNN) and Decision Tree (DT). With recommendations from [24], after the implementation of several machine learning techniques, ensemble methods will be applied to improve prediction accuracy. Machine learning techniques from the previous section will be combined for bagging, boosting and stacking with the hyperparameter tunings determined from the base models.

D. Performance Assessment

In this stage of the study, K-fold cross-validation, a popular performance evaluation technique will be applied. The dataset is divided into K equal-sized subsets, or "folds", and the model is then trained and evaluated K times, using a different fold as the test set and the remaining K-1 fold as the training set each time. Every data point is used for both training and testing thanks to this procedure. The performance measure is noted following the use of each fold as the test set. A single estimate of the model's performance is then generated by averaging the K findings. Since the split value was set to 5 in the code implementation and the number of rows in the dataset is equivalent to 254745 rows, this means that when creating subsets, 80% ends up being the training set whilst 20% ends up being the testing set for each iteration. In other words, a 80-20 split happens at every iteration. By training the model several times on various subsets of the data, K-fold cross validation minimises overfitting and offers a more reliable assessment than a single train-test split. It contributes to the model's ability to function well when applied to unknown data. When a model's performance can vary greatly with different training sets, the average performance across folds provides a more accurate estimate of the model's actual performance.

IV. RESULT & DISCUSSION

Table 2 depicts the performance metrics of the base models. It is observed that KNN achieved the highest accuracy of 85.904% and it had a precision of 71.05%, recall of 56.38% and f1-scores of 57.789%. In the decreasing order of accuracy, DT was next at 85.89%, followed by SVM at 85.52% and lastly was NB, with the lowest accuracy of 83.039%.

Table 2: Performance Metrics of Base Models

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Scores (%)
KNN	85.904	71.050	56.380	57.789
DT	85.888	70.754	57.211	58.940
NB	83.039	65.627	65.396	65.506
SVM	85.518	42.759	50.000	46.097

For ensemble methods, bagging in particular, it is observed from Table 3 that the top performing model is RF which has the highest accuracy with the value of 85.903%, a precision of 71.019%, recall of 56.486% and F1-score of 57.939%. its outperformance is shown through its confusion matrix in fig. 5. It is then followed by Bagged-DT which has the second highest accuracy of 85.901%, Bagged-KNN with an accuracy of 85.885%, Bagged-SVM at 85.518% and lastly Bagged-NB with the lowest accuracy at 85.027%.

Table 3: Performance Metrics of Bagging Methods

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Scores (%)
RF	85.903	71.019	56.486	57.939
Bagged - KNN	85.885	70.901	56.489	57.939
Bagged - DT	85.901	70.988	56.513	57.976
Bagged - NB	85.027	66.239	56.453	57.793
Bagged - SVM	85.518	42.759	50.000	46.097

For ensemble methods, boosting in particular, it is observed from Table 4 that GB had the highest accuracy with the value 85.904% along with precision, recall and f1-scores values of 70.921%, 56.840% and 58.439% respectively. Following in decreasing order of accuracy was XGB at 85.902% and AB at 85.822% which was the lowest accuracy for boosting.

Table 4: Performance Metrics of Boosting Methods

Models	Accuracy (%)	Precision (%)	Recall (%)	F1-Scores (%)
AB	85.822	70.385	56.869	58.468
GB	85.904	70.921	56.840	58.439
XGB	85.902	71.049	56.374	57.780

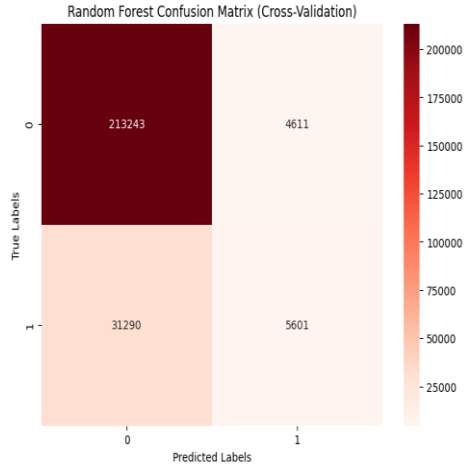

Fig. 5: Confusion matrix for RF (Bagging)

From the overall results, we can clearly see that all the models, base and ensemble, were highly competitive with a high majority of results achieving approximately an 85% accuracy. The results were very close to the point where comparing the results to the nearest 3 decimal places was needed. Out of all the ensembled methods used, the best performing ensemble method was boosting with Gradient Boosting (GB) giving an accuracy of 85.904%, precision of 70.921%, recall of 56.840% and f1-score of 58.439%.

Despite GB achieving the highest accuracy among the ensemble methods though, it is observed that GB actually has the same accuracy as the base model K-Nearest Neighbours (KNN) which is also 85.904% but with different precision, recall and f1-scores instead at, 71.050%, 56.380% and 57.789% respectively. While both models have similar accuracy, GB has a slight edge over KNN due to its higher recall and F1-score, indicating a better balance and effectiveness in identifying true positives while managing false positives effectively, as shown in fig. 6. Therefore, GB is considered as the better model overall which makes it the best model out of the base and ensemble techniques used. The comparable performance of KNN in specific, to other ensemble techniques does raise the question about the added value of the extensive ensemble models used in this project. This indicates that the dataset's simplicity of lack of data balancing may have limited the added value of hybrid methods. Ensemble methods, which are designed to capture complex patterns, often excel with balanced datasets, while KNN's straightforward approach can perform well when relationships in the data are relatively simple. Additionally, effective preprocessing, such as RFECV, might have reduced redundancies, enabling simpler models like KNN to perform competitively. However, metrics like F1-score and recall as mentioned in the result discussion, does show the weaknesses of this study in handling the imbalanced data, which likely hindered the full potential of ensemble methods. If the data imbalance was addressed in the first place, maybe the evaluation process would've demonstrated a better value of hybrid approaches.

This study presents a novel comparative analysis of ensemble machine learning techniques for diabetes prediction, integrating multiple hybrid approaches bagging, boosting, and stacking to evaluate their effectiveness. Unlike previous studies that focus on isolated algorithms, this research systematically investigates how different ensemble methods enhance classification accuracy, precision, recall, and F1-score. Additionally, the use of Recursive Feature Elimination with Cross-Validation (RFECV) for feature selection and K-Fold Cross-Validation for model evaluation ensures robust performance assessment. By leveraging an extensive dataset from the Behavioral Risk Factor Surveillance System, this work provides new insights into the practical implementation of machine learning in healthcare, particularly for early diabetes diagnosis, paving the way for improved clinical decision-making.

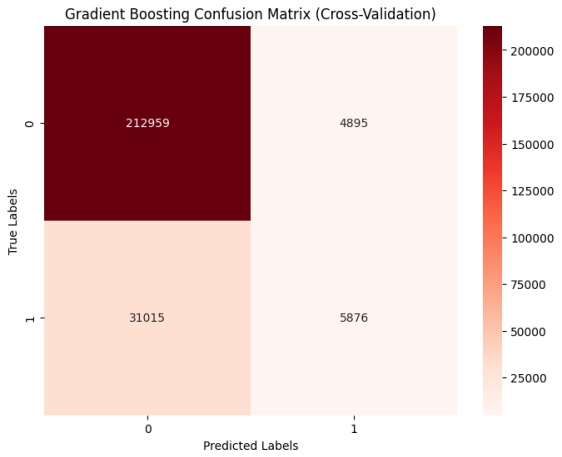

Fig. 6: Confusion matrix for GB (Boosting)

V. CONCLUSION

To conclude this study, it is rather important to continually develop techniques in the machine learning realm in order to come up with solutions that can help improve the quality of life of diabetic patients by providing early intervention through predictive methods. This study managed to propose and carry out an effective framework for diabetes prediction and got decent results from it. Many recommendations were taken from several journals such as to carry out feature selection and so on (Refer section 2.7), which allowed the improvement of results majorly compared to other studies that used the same dataset (2015 survey on Behavioral Risk Factor Surveillance System for diabetes prediction) to predict diabetes. Additionally, thorough actions of grid search for hyperparameter tuning may have also contributed to the improved results. In particular, the boosting method, Gradient Boosting managed in the end to outperform the other methods, by providing an accuracy of 85.904% where, despite being the same as KNN, was still considered to be better due to its higher recall of 56.840% and f1-score of 58.439%. For other ensemble methods such as bagging, RF got the highest accuracy of 85.903% and for stacking, DT+SVM managed to get the highest accuracy of 85.87%. Limitations that can be seen from this study would probably be the involvement of few base algorithms only (DT, KNN, SVM, NB) and the inability to conduct hyperparameter tuning with grid search on SVM due to computational limitations. Future improvements should involve more suitable base algorithms to the process as well as more thorough

hyperparameter tuning to all algorithms even the ensemble ones instead of just using the optimal parameters from its base model or default parameters itself.

REFERENCES

[1] Yachmaneni Jr, A., et al., A comprehensive review of the vascular consequences of diabetes in the lower extremities: current approaches to management and evaluation of clinical outcomes. Cureus, 2023. **15**(10).

[2] Herman, S.T., et al., Consensus statement on continuous EEG in critically ill adults and children, part I: indications. Journal of Clinical Neurophysiology, 2015. **32**(2): p. 87-95.

[3] Tsao, C.W., et al., Heart disease and stroke statistics—2023 update: a report from the American Heart Association. Circulation, 2023. **147**(8): p. e93-e621.

[4] Kulkarni, A., A.R. Thool, and S. Daigavane, Understanding the Clinical Relationship Between Diabetic Retinopathy, Nephropathy, and Neuropathy: A Comprehensive Review. Cureus, 2024. **16**(3).

[5] Ajibade, S.-S.M., et al., Evolution of machine learning applications in medical and healthcare analytics research: A bibliometric analysis. Intelligent Systems with Applications, 2024: p. 200441.

[6] Kong, L.S., et al., A systematic review on software reliability prediction via swarm intelligence algorithms. Journal of King Saud University-Computer and Information Sciences, 2024: p. 102132.

[7] Ajibade, S.-S.M., et al. Enhancing Students' Learning Motivation and Comprehension by Reflecting on the Practical Applications of Learning Materials in an Education Learning Journal. in 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG). 2024. IEEE.

[8] Vaidya, S., et al., A computer-aided feature-based encryption model with concealed access structure for medical Internet of Things. Decision Analytics Journal, 2023. **7**: p. 100257.

[9] Dou, Q., J. Zhang, and B. Jing, A ML-based economic protection development level using Decision Tree and Ensemble Algorithms. Soft Computing, 2023. **27**(24): p. 18929-18947.

[10] Ajibade, S.-S.M., et al. Application of Artificial Intelligence in Healthcare Systems: A Scientometric Analysis. in 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG). 2024. IEEE.

[11] Freschet, G.T., et al., A starting guide to root ecology: strengthening ecological concepts and standardising root classification, sampling, processing and trait measurements. New Phytologist, 2021. **232**(3): p. 973-1122.

[12] Ghosh, S., A. Dasgupta, and A. Swetapadma. A study on support vector machine based linear and non-linear pattern classification. in 2019 International Conference on Intelligent Sustainable Systems (ICISS). 2019. IEEE.

[13] Saravanan, K., et al. Support Vector Machines: Unveiling the Power and Versatility of SVMs in Modern Machine Learning. in 2023 3rd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA). 2023. IEEE.

[14] Goel, A. and S.K. Srivastava. Role of kernel parameters in performance evaluation of SVM. in 2016 Second international conference on computational intelligence & communication technology (CICT). 2016. IEEE.

[15] Jovin, T.G., et al., Trial of thrombectomy 6 to 24 hours after stroke due to basilar-artery occlusion. New England Journal of Medicine, 2022. **387**(15): p. 1373-1384.

[16] Adetunla, A., et al. Analysing the Roles of Robotics in Manufacturing Organizations in the Era of Industry 4.0. in 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG). 2024. IEEE.

[17] Zaidi, A., et al., Evolution of climate-smart agriculture research: A science mapping exploration and network analysis. Open Agriculture, 2024. **9**(1): p. 20220396.

[18] Ajibade, T.I., et al. Statistical Analysis of Digital Financial Technology Adoption Research. in 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG). 2024. IEEE.

[19] Nagesh, O.S., et al., Boosting enabled efficient machine learning technique for accurate prediction of crop yield towards precision agriculture. Discover Sustainability, 2024. **5**(1): p. 78.

[20] Yap, J.J., et al. Improving Object Detection in Videos: A Comprehensive Evaluation of Faster R-CNN Employed in Partial Occlusion Handling. in 2024 20th IEEE International Colloquium on Signal Processing & Its Applications (CSPA). 2024. IEEE.

[21] Saihood, Q. and E. Sonuç, A practical framework for early detection of diabetes using ensemble machine learning models. Turkish Journal of Electrical Engineering and Computer Sciences, 2023. **31**(4): p. 722-738.

[22] Lee, C.H. and H.-J. Yoon, Medical big data: promise and challenges. Kidney research and clinical practice, 2017. **36**(1): p. 3.

[23] Tan, L.Y., et al. Artificial Intelligence Models in Power Generation for Energy Consumption Prediction. in 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC). 2024. IEEE.

[24] Chweya, R., S.-S.M. Ajibade, and A.J. Melbury, The importance and limitations of big data technologies in education, in Recent Advances in Material, Manufacturing, and Machine Learning. 2023, CRC Press. p. 1449-1454.