ISSN 2723-6471

A Comprehensive Stacking Ensemble Approach for Stress Level Classification in Higher Education

Hendry Fonda[1,*], Yuda Irawan[2], Rika Melyanti[3], Refni Wahyuni[4], Abdi Muhaimin[5]

[1,3,5]*Information System, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

[2,4]*Computer Science, Universitas Hang Tuah Pekanbaru, Pekanbaru, Indonesia*

(Received: August 23, 2024; Revised: September 14, 2024; Accepted: October 05, 2024; Available online: October 15, 2024)

Abstract

This research focuses on developing a comprehensive stacking ensemble model for the classification of student stress levels in higher education environments, specifically at Hang Tuah University Pekanbaru. Using a physiological dataset that includes parameters such as SPO2, heart rate, body temperature, systolic, and diastolic pressure, this research categorizes the condition of college students into four main categories: anxious, calm, tense, and relaxed. Data from 2021 to 2024 was processed using the SMOTE technique to address data imbalance, and K-Fold Cross Validation (K=10) was applied for robust model validation. In model development, a combination of basic algorithms such as SVM, Logistic Regression, Multilayer Perceptron, and Random Forest is used which is enhanced by boosting techniques through ADABoost, and XGBoost as a meta model. The test results show that the proposed stacking model is able to achieve 95% accuracy, with an AUC of 0.95, which indicates excellent performance in classification. The model not only excels in detecting more extreme stress conditions such as anxiety, but also shows reliable ability in classifying more difficult to distinguish conditions such as tense and relaxed. The conclusion of this study shows that the applied stacking ensemble approach significantly improves prediction accuracy and stability compared to traditional models. For future research, it is recommended to explore the use of deep learning-based meta-models such as LSTM and BiLSTM as well as rotation techniques in stacking to improve model performance and flexibility. The findings are expected to contribute significantly to the development of more sophisticated and effective stress detection models.

Keywords: Stacking, Boosting, SMOTE, Machine Learning, Stress Level Classification

1. Introduction

In an increasingly demanding era of higher education, students are faced with various academic and emotional challenges that can affect their academic performance. One significant challenge that is often overlooked is the level of stress experienced by students, particularly during the thesis completion process. Poorly managed stress can negatively impact students' ability to focus, make decisions, and complete their academic tasks. At Hang Tuah University Pekanbaru Faculty of Computer Science, it was noted that only 75% of students managed to complete their thesis on time, which indicates a problem that needs to be addressed. One of the main causes of this low on-time completion rate is undetected and poorly managed stress by academic advisors, who often do not have the right tools or indicators to identify stress levels in students.

The stress experienced by students comes not only from academic pressure, but also from physiological factors related to their physical health. Parameters such as SPO2, heart rate, temperature (body temperature), blood pressure systolic and diastolic can provide an overview of the physical and emotional condition of students [1]. With the advancement of wearable technology, physiological data collection has become easier and more accurate [2], [3]. However, in many educational institutions, monitoring of these parameters has not been an integral part of the academic advising process. This leads to a gap in early detection of stress in students, which in turn affects the effectiveness of thesis guidance and timely completion of studies. The expected outcomes of this research include the development of a hybrid machine learning model that is able to detect students' stress levels and categorize them into four main categories: anxious, calm, tense, and relaxed. With this model, it is expected to help the study program in monitoring students' conditions more comprehensively, so that the necessary interventions can be carried out timelier and effectively. This categorization

***Corresponding author: Hendry Fonda (fondaanda@gmail.com)**
DOI: https://doi.org/10.47738/jads.v5i4.388

will not only provide a deeper insight into students' emotional state, but also allow academic advisors to provide more personalized and contextualized support, which in turn can increase the percentage of on-time graduation.

The ensemble stacking technique has proven to be an effective method for improving prediction accuracy in various domains, including disease detection and medical event prediction [4]. In a study by Velmurugan and Dhinakaran, they proposed an ensemble stacking method that combines algorithms such as Random Forest (RF), XGBoost, and multilayer perceptron (MLP) for Parkinson's disease prediction. They found that this approach significantly improved the prediction accuracy compared to other single methods [5]. In addition, a study by Khairul Islam et al. developed a stacking model that integrates machine learning algorithms to predict emergency revisit rates in heart disease patients, showing that this hybrid model is able to produce predictions with high accuracy as well as improve the generalization ability of the model [6]. In a related study by Merdassi et al., a stacking-based model was also used in drought forecasting, where they showed that the stacking model provides more stable and accurate prediction results in a changing environment [7]. Another study developed a stacked ensemble-based PSVM-PMLP-MLR hybrid model to predict energy consumption in the electrolytic copper foil manufacturing process. The results show that this model is able to improve prediction accuracy by reducing the absolute mean error (MAE) and increasing the regression coefficient (R^2) compared to single models such as SVM and MLP [8]. In another study, a stacking ensemble method was proposed to detect three types of diabetes mellitus using a dataset from Saudi Arabia. This method showed significant improvement in detection accuracy and prediction stability, compared to other ensemble techniques such as bagging [9].

In addition, the study used the stacking ensemble method to predict energy consumption in metro systems with better results in overcoming non-linearity and variable interaction problems compared to traditional models [10]. Researchers further developed a novel stacking technique for diabetes prediction using the PIMA Indian dataset, where they combined MLP, support vector machine (SVM), and logistic regression (LR) as base models. This technique showed improved accuracy compared to other methods such as AdaBoost [11]. Other researchers have shown significant improvements in the prediction of major cardiovascular events [12]. In addition, it uses a stacking-based algorithm for social phobia classification, which also shows the superiority of this technique in dealing with the complexity of psychological data [13]. A stacked ensemble technique for fireproof column classification that combines various machine learning algorithms to improve classification accuracy [14]. Another researcher developed a stacked ensemble model for type 2 diabetes prediction using a combination of algorithms such as KNN, SVM, RF, and Naive Bayes as base models, with Logistic Regression as a meta-model, which gave a prediction accuracy of 94.17% [15].

This research offers several advantages over previous studies with ensemble stacking optimization by combining powerful base models such as SVM, Logistic Regression, Multilayer Perceptron, and Random Forest, each of which is enhanced with boosting techniques through ADABoost, this research seeks to significantly improve prediction accuracy. The use of XGBoost as a meta model adds a better predictive layer to optimize the prediction results of the existing base models. In addition, this study uses a physiological dataset that includes parameters such as SPO2, heart rate, body temperature, systolic and diastolic pressure, taken from public health centers in Riau Province, Indonesia. This dataset was processed with SMOTE technique to handle data imbalance, and K-Fold validation for 10 times to ensure that the resulting model is not only accurate but also robust in the face of data variation. This research also takes a case study of students from the Faculty of Computer Science, Hang Tuah University Pekanbaru, which specifically focuses on detecting students' stress levels in the context of completing academic tasks such as thesis. With this comprehensive approach, this research is expected to provide a significant improvement in the accuracy of student stress detection compared to existing methods, as well as provide a deeper insight into the physiological conditions that contribute to stress in an academic context.

2. Research Methodology

This research method focuses on developing a staking model with the application of various machine learning techniques to detect student stress levels based on physiological data. The research process involves several important stages that ensure the resulting model has high accuracy. The stages of model development can be seen in figure 1 below:

 ISSN 2723-6471

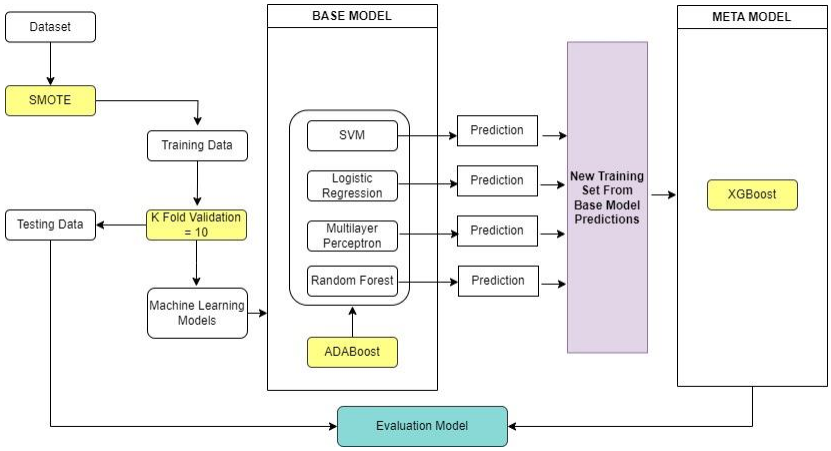

Figure 1. Development of Stacking Model

2.1. Dataset

The physiological dataset used in this study was collected from public health centers (Puskesmas) in Riau Province, Indonesia, between 2021 and 2024. The data includes vital physiological parameters such as SPO2, heart rate, body temperature, systolic, and diastolic blood pressure, with a total of 2,851 entries. These physiological metrics were recorded during routine medical check-ups of university students, specifically during academic stress periods, such as exam preparations and thesis completion. Data collection was conducted using standardized, medically certified instruments: pulse oximeters for SPO2 and heart rate, digital thermometers for body temperature, and digital blood pressure monitors for systolic and diastolic pressure. All instruments were calibrated before data collection to ensure reliability and accuracy. This dataset is designed to monitor and analyze the physiological state of college students in four main categories: Anxious, Calm, Tense, and Relaxed. This data aims to provide insight into the stress levels of students during the process of completing academic tasks, such as thesis. The dataset view is seen in table 1 below:

Table 1. Fisiologis Dataset

SPO2	Heart Rate	Temperature	Systolic	Diastolic	Condition
96	96	34,4	125	95	Anxious
97	100	34,0	120	98	Anxious
91	88	35,6	113	85	Calm
95	82	36,0	118	88	Calm
93	78	35,6	111	89	Calm
98	85	34,7	120	96	Anxious
100	90	34,7	124	100	Anxious
104	105	34,4	141	120	Tense
97	100	34,9	129	99	Anxious

Preprocessing of the physiological dataset was carried out to ensure high-quality data for model training and evaluation. The steps included handling missing values and detecting outliers. Missing data were addressed by imputing the mean value for each respective column, as the missing values comprised less than 5% of the dataset, making mean imputation an effective and unbiased solution. Outliers were detected using the Interquartile Range (IQR) method, where values falling below Q1 - 1.5IQR or above Q3 + 1.5IQR were considered outliers. These outliers were either corrected (in cases of measurement errors based on expert advice) or removed if they represented rare physiological anomalies that could distort model learning. Additionally, for continuous variables like heart rate and blood pressure, the Z-score method was used as a supplementary check to identify extreme outliers, with values beyond ±3 flagged for further treatment. This systematic approach ensures data integrity and enhances the reproducibility of the study.

Ethical measures were implemented to ensure the responsible handling of sensitive health data. Participants provided informed consent, understanding the study's purpose and their right to withdraw. The dataset was anonymized by

removing personal identifiers and assigning unique codes to protect privacy. Additionally, the data was stored securely on a server accessible only to authorized researchers. These safeguards ensured that participants' rights and privacy were protected throughout the study.

2.2. SMOTE

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to handle data imbalance in machine learning [16], [17]. Data imbalance occurs when the number of samples in one class is much less than the other classes, which can result in predictive models tending to ignore the minority class. SMOTE works by creating synthetic samples of the minority class by interpolating between existing samples and simply duplicating existing data [18]. This is done by selecting adjacent points in the feature space and creating new data based on a linear combination of these points. By using SMOTE, the class distribution in the dataset becomes more balanced, allowing the model to learn better and produce more accurate predictions, especially in recognizing patterns from previously underrepresented minority classes [19]. To demonstrate the effectiveness of SMOTE, a comparative analysis of the dataset before and after applying SMOTE is shown in table 2 below:

Table 2. Comparative Analysis of The Dataset Before and After Applying SMOTE

Stress Level	Before SMOTE	After SMOTE
Anxious	1200	1200
Calm	850	1200
Tense	500	1200
Relaxe	301	1200

As shown in the table 1, SMOTE balanced the number of instances in each class, ensuring that the model received an equal representation of data from all categories. This adjustment was crucial in enhancing the model’s ability to learn from the underrepresented classes, thereby improving its overall classification performance.

2.3. K-Vold Cross Validation

K-Fold Cross Validation is a technique used to more accurately evaluate the performance of machine learning models by dividing the dataset into multiple subsets or folds [20]. In this study, K-Fold Cross Validation was used with a value of K=10, which means that the dataset of 2,851 rows was divided into 10 subsets of similar size. At each iteration, one of the subsets was used as test data, while the other nine subsets were used as training data. This process was repeated 10 times, with each subset being used as test data once. The decision to use K=10 was based on its balance between bias and variance, computational efficiency, and empirical support. K=10 is widely regarded as a good compromise, providing reliable results by reducing high variance seen with smaller K values and minimizing bias found with larger K values. It also offers efficient model validation without excessive computational cost, particularly for datasets of moderate size like ours. Empirical evidence from machine learning studies consistently demonstrates that K=10 produces stable, robust results, making it an optimal choice for cross-validation in this study. K-Fold Cross Validation ensures that the model is trained and tested on the entire dataset, resulting in more stable evaluation results and better generalization of the model to unseen data [21], [22]. By applying this technique to the physiological dataset used, this research aims to minimize the risk of overfitting and provide a more reliable assessment of the model's performance in detecting college students' stress levels.

2.4. Stacking Ensemble Model

Stacking is an ensemble technique in machine learning that combines predictions from multiple base models by using meta models to produce more accurate final predictions [23]. Table 3 of the literature review below summarizes previous studies that used stacking techniques in the development of machine learning models:

Table 3. The previous research related to stacking

Researcher	Based Model	Meta Model	Accuracy
Rezaei Melal [24]	KNN, Decision Tree, RF, XGBoost	Neural Network (NN)	94.0%
Nyaramneni [25]	RF, XGBoost, LGBM	LR	94.7%

 ISSN 2723-6471

Almohimeed [26]	RF, DT, SVM, LR, KNN, NB	RF	90.03
Qian-Chuan [27]	KNN, RF, Support Vector Regression (SVR)	RF	93.8%
Kshatri [28]	SVM, J48, Naïve Bayes, Bagging, Random Forest	SVM	94.5%
Seireg [29]	LGBM, GBR, XGBoost	Ridge Regression	93.5%

Previous studies listed in table 3 show a variety of approaches in applying stacking techniques to improve prediction accuracy. Some studies such as those by Rezaei Melal and Nvarameni used a combination of base models such as KNN, Decision Tree, Random Forest, and XGBoost, with meta-models such as Neural Network and Logistic Regression, which resulted in high accuracy of 96.0% and 94.7%, respectively. Other studies, such as by Almohiweed, also explored the use of various base models and meta-models, including the application of RF and AdaBoost algorithms, to achieve accuracy of up to 93.8%. Overall, these studies show that the use of stacking techniques with the right combination of base models and meta-models can significantly improve predictive performance, with accuracy varying from 90.03% to 96.0%. Table 4 shows the proposed development of stacking models that are expected to improve accuracy.

Table 4. Proposed Stacking Model

Variable	Optimization	Based Model	Meta Model
SP02, Heart Rate, Temperature, Systolic, Diastolic	SMOTE K-Fold Cross Validation Boosting (ADABoost)	SVM LR MLP RF	XGBoost

The stacking model proposed in table 4 shows significant advantages over the previous studies summarized in table 2. The model incorporates several optimization techniques such as SMOTE, K-Fold Cross Validation, and Boosting to ensure data quality and balance and improve prediction performance.

The stacking process begins with the independent training of each base model, utilizing algorithms such as SVM, LR, MLP, and RF. Each of these models is trained separately on the preprocessed physiological data, where they learn to predict one of the four stress categories: Anxious, Calm, Tense, and Relaxed. After the base models are trained, their predictions, represented as class probabilities for each stress category, are used as input features for the meta model. Specifically, for each instance in the dataset, the base models generate probability distributions over the four categories, which are then combined into a single feature vector. This feature vector serves as a comprehensive representation of the collective knowledge gathered from all base models. The meta model, implemented using XGBoost, is then trained using these feature vectors. XGBoost was chosen due to its ability to effectively optimize decision trees and its strong performance in handling overfitting. The meta model takes the outputs from the base models as inputs and learns how to combine them in an optimal way to produce the final predictions. During the training process, it assigns appropriate weights to the predictions of the base models based on their individual performance, leading to a more accurate and reliable prediction of the stress categories. This stacked approach leverages the strengths of each base model and enhances overall predictive performance through a hierarchical learning strategy.

The meta model automatically determines how much weight to assign to each base model's predictions based on their contributions to improving the final classification accuracy. Base models that perform better on certain stress categories are given higher weights for those categories. For example, Random Forest, which performed particularly well in predicting the "Anxious" class, was given a higher weight for that class. XGBoost, with its boosting capability, ensures that any misclassifications made by the base models are corrected, thereby improving the overall accuracy of the ensemble. This holistic approach, which utilizes the strengths of each algorithm in handling data variation and imbalance, is expected to provide superior results compared to the models used in previous studies.

2.5. Model Evaluation

Model evaluation in this study was conducted using two main metrics Confusion Matrix and Receiver Operating Characteristic (ROC) curve. Confusion Matrix provides a detailed overview of the model's performance by showing the number of correct and incorrect predictions for each class, including True Positives, False Positives, True Negatives,

 ISSN 2723-6471

and False Negatives [30], [31]. This allows for in-depth analysis of classification errors, especially in the context of data imbalance. Meanwhile, the ROC curve is used to evaluate the model's ability to distinguish between positive and negative classes at various thresholds, with the Area Under the Curve (AUC) as an indicator of overall model performance [32], [33]. An AUC close to 1 indicates a model with excellent performance in prediction. By using these two metrics, the model evaluation becomes more comprehensive, ensuring that the model is not only accurate overall, but also able to predict well for each class it tests.

3. Result and Discussion

A dataset obtained from public health centers in Riau has been analyzed in depth to obtain important information regarding the physiological conditions of university students. The dataset includes 2,851 data collected over the period 2021 to 2024. The data includes key physiological parameters such as SPO2, Heart Rate, Temperature, Systolic and Diastolic, which are categorized into four physiological states: Anxious, Calm, Tense, and Relaxed. This data can also be collected using IoT devices, for example to measure body temperature [34]. The following correlation heatmap depicts the relationship between various physiological variables, namely SPO2, Heart Rate, Temperature, Systolic, and Diastolic. Figure 2 below shows the Heatmap Correlation Matrix of physiological parameters:

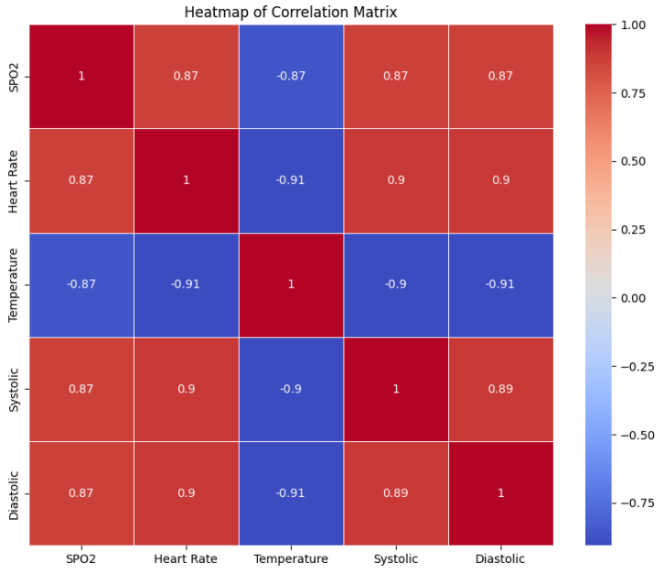

Figure 2. Heatmap Correlation Matrix

The matrix image above shows correlation values ranging from -1 to 1, where positive values indicate a unidirectional relationship (when one variable increases, the other variable also increases), and negative values indicate an opposite relationship (when one variable increases, the other variable decreases). From this heatmap, it can be seen that there is a strong positive correlation between Heart Rate, Systolic, and Diastolic, with correlation values of around 0.9 each. This indicates that when systolic or diastolic blood pressure increases, heart rate tends to increase as well. In contrast, Temperature showed a strong negative correlation with SPO2 and Heart Rate, with correlation values around -0.87 to -0.91, meaning that an increase in body temperature tends to be followed by a decrease in oxygen saturation and heart rate.

The distribution of these physiological conditions shows the variation in how students experience stress, with a particular focus on the impact these conditions have on their academic performance, especially when completing final assignments. Here is figure 3 of the results of labeling the conditions of students' stress levels. Figure 3 illustrates the distribution of these physiological states across the student population, showing significant variation in the prevalence of each state. The graph of the distribution of the physiological states of the college students shows a significant imbalance of data, with the condition “Anxious” having a significantly larger amount of data compared to other states such as “Relaxed” and “Tense.” This imbalance may cause machine learning models to be more accurate in predicting the more dominant state (Anxious) and less effective in detecting states with less data. To address this issue, the

 ISSN 2723-6471

SMOTE technique is used. SMOTE synthetically generates new samples for minority classes by creating a linear combination of existing data, so that the data distribution becomes more balanced. With the application of SMOTE, the model is expected to learn better from each category of physiological conditions, thus improving the accuracy and generalization of the prediction. The following figure 4 displays a graph of the balanced data after SMOTE.

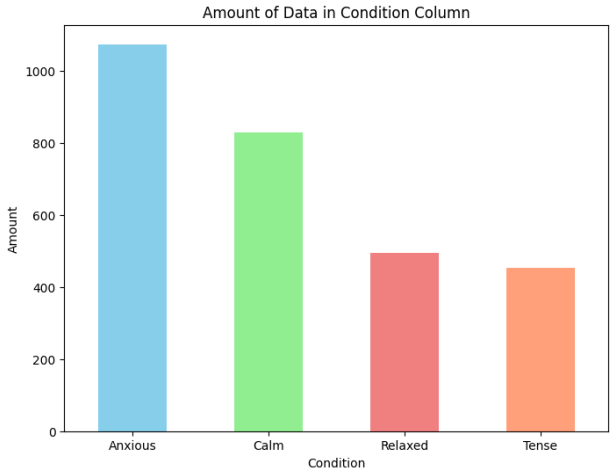

Figure 3. The Real Label Graph

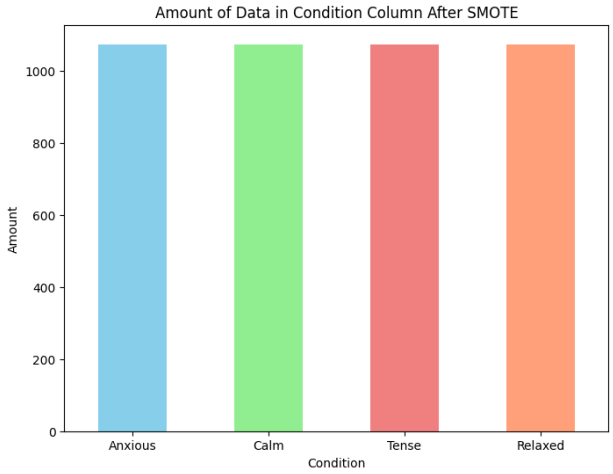

Figure 4. The Label Graph After SMOTE

After the application of SMOTE, the data distribution in each state category (Anxious, Calm, Tense, and Relaxed) became balanced, with almost the same amount of data in each category, which is about 1000 samples. This balance indicates that SMOTE successfully overcomes the problem of initial data imbalance, so that the model can be trained with a more even representation of each condition, which is expected to improve the accuracy and generalization of the model. The following table 5 compares the accuracy of the algorithm on the base model with the dataset before and after SMOTE:

Table 5. Comparison of accuracy of SMOTE application on Base Model

Base Model Algorithm	Accuracy	
	Without SMOTE	With SMOTE
SVM	82%	85%
LR	83%	82%
MLP	82%	84%
RF	84%	87%

Table 5 above shows the accuracy comparison of several machine learning algorithms on the base model before and after applying SMOTE. From the results shown, it can be seen that the application of SMOTE succeeded in improving the accuracy of most algorithms. For example, the SVM algorithm experienced an increase in accuracy from 82% to 85%, and RF increased from 84% to 87%. This shows that SMOTE is effective in handling data imbalance problems, so that the model can learn better and provide more accurate predictions. However, in LR, the accuracy slightly decreased from 83% to 82%, which may be due to overfitting or mismatching of the algorithm with the offset data. Overall, SMOTE was shown to improve model performance on imbalanced data, especially on algorithms such as SVM, MLP, and RF.

The application of ADABoost to each base model in this study aims to improve the accuracy and strengthen the predictive ability of each base model, namely SVM, Logistic Regression, MLP, and Random Forest. ADABoost works by building a series of weak learners iteratively, where each new model is built to correct the errors of the previous model. The final result is a combination of all models that pay attention to the weight of each prediction based on the accuracy of each model. The analysis in table 4 shows that the application of ADABoost is effective in improving the overall accuracy of the base model, especially for models such as SVM and Random Forest that have shown high performance before. In base models such as Logistic Regression and MLP, ADABoost helps reduce the variation in

 ISSN 2723-6471

accuracy seen during cross-validation, providing more stable and reliable prediction results. Table 6 shows a comparison of the accuracy of applying ADABoost to the base model:

Table 6. Comparison of the accuracy of applying ADABoost to the Base Model

Base Model Algorithm	Accuracy	
	Without ADABoost	With ADABoost
SVM	85%	88%
LR	82%	85%
MLP	84%	87%
RF	87%	89%

The graphs in table 6 show the accuracy comparison between the base models before and after the application of ADABoost. From the results shown, it can be seen that ADABoost consistently improves the accuracy of all base models, with the improvement varying between 2% to 3%. For example, SVM accuracy increased from 85% to 88%, and Random Forest from 87% to 89%. These improvements show that ADABoost is effective in improving the model's ability to classify data, making it more accurate and reliable. In this study, K-Fold Cross Validation with a value of K=10 was applied to ensure that the machine learning model used was able to produce consistent and reliable performance. This technique divides the dataset into 10 subsets, where each subset in turn is used as test data, while the other nine subsets are used as training data. This process is repeated 10 times, allowing for a comprehensive evaluation of the model's accuracy. The results of K-Fold Cross Validation show a stable distribution of accuracy, which is displayed in the form of a boxplot, ensuring that the model is not only accurate overall but also able to generalize well to data that has never been seen before. Figure 5 is a visual of the application of this technique providing strong evidence that the resulting model is resilient to data variations, making it reliable in different situations.

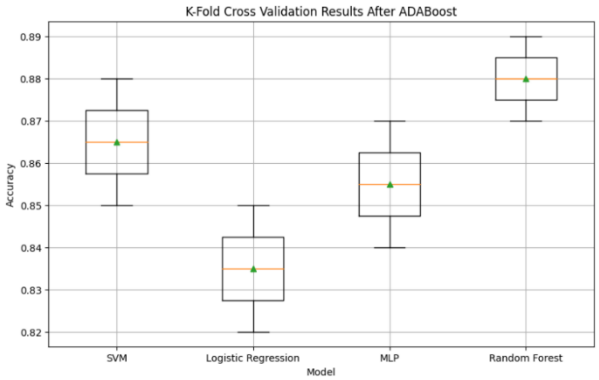

Figure 5. K-Fold Cross Validation After ADABoost

The boxplot graph above displays the K-Fold Cross Validation results after the application of ADABoost. From this graph, it can be seen that Random Forest has the highest accuracy with a relatively narrow distribution, indicating that this model is consistent in providing good performance after the application of ADABoost. SVM also shows a significant increase in accuracy with lower variability, indicating good stability. On the other hand, Logistic Regression has a wider accuracy distribution and lower mean value than the other models, indicating that although ADABoost improves performance, it is still less stable. Overall, the application of ADABoost successfully improved the accuracy and consistency of performance of most models, especially Random Forest and SVM. After the data balancing process using SMOTE and boosting using ADABoost, the next step is to carry out the classification process using the stacking ensemble model. As an algorithm based on ADABoost (SVM, LR, MLP, and RF) with XGBoost as a meta model. Table 7 is a classification report from testing the stacking model with the XGBoost meta model.

Table 7. Classification Report in Stacking Ensemble

	Precision	Recall	F1-score	Support
Anxious	0.93	0.96	0.94	800

 ISSN 2723-6471

Calm	0.94	0.93	0.93	750
Tense	0.92	0.90	0.91	650
Relaxed	0.91	0.94	0.92	651
Accuracy			0.95	2851
Macro avg	0.93	0.93	0.93	2851
Weight avg	0.94	0.95	0.94	2851

Table 7 shows the classification report of the test results of this model, showing that the stacking model is able to achieve an overall accuracy of 95%. The Precision, Recall, and F1-score values for each condition category (Anxious, Calm, Tense, and Relaxed) are quite consistent, with the highest Precision and Recall in the "Anxious" category reaching 0.93 and 0.96. Although the "Tense" condition has a slightly lower Recall value, namely 0.90, this model still provides good overall performance, indicated by the high macro average and weighted average values (0.93 and 0.94, respectively). This confirms that the combination of SMOTE, ADABoost (Base Model), and XGBoost as Meta Models in this stacking model is effective in handling imbalanced data and providing accurate and reliable classification results. Then figure 6 is the result of the confusion matrix.

In figure 6, the Confusion Matrix above shows the model's performance in classifying four physiological conditions: Anxious, Calm, Tense, and Relaxed. Overall, the model performed well in detecting the "Anxious" condition with 760 correct predictions and only 20 misclassified as "Calm," 10 as "Tense," and 10 as "Relaxed." In the "Calm" condition, the model also performed quite well with 700 correct predictions, but there were still 30 misclassifications to "Anxious," and 10 each to "Tense" and "Relaxed." For the "Tense" condition, the model successfully identified 590 cases correctly, but there were 20 misclassifications to "Anxious," 20 to "Calm," and 20 to "Relaxed." While for the "Relaxed" condition, the model produced 611 correct predictions, but there were 15 mistakes to "Anxious," 10 to "Calm," and 15 to "Tense". Misclassification occurs mainly in conditions that have similar physiological features, such as "Tense" and "Relaxed," which are difficult for the model to distinguish. Although the model is quite reliable in detecting more extreme conditions, such as "Anxious," further improvements are needed to reduce misclassification between more difficult-to-distinguish categories, in order to improve overall accuracy. To improve the model's differentiation between "Tense" and "Relaxed" states, several strategies can be explored. Feature engineering could introduce measures like heart rate variability (HRV) and time-series data to capture more subtle physiological changes, with models like LSTM applied to recognize temporal patterns. Class-specific tuning and cost-sensitive learning can help address classification challenges by penalizing misclassifications more heavily. Additionally, hierarchical classification could first categorize broader stress levels before distinguishing between similar states. Figure 7 shows the results of testing using ROC.

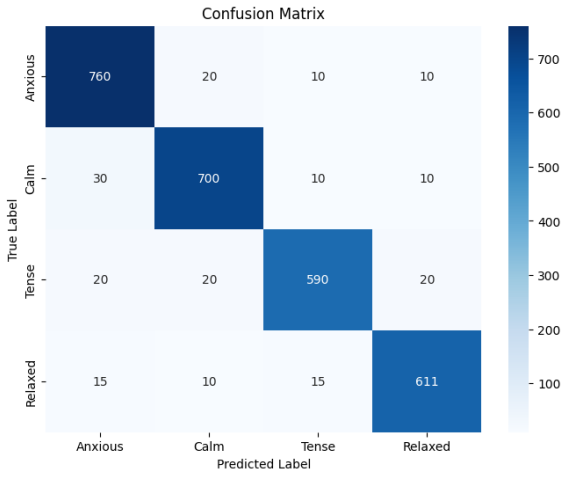

Figure 6. Test Results Using Confusion Matrix

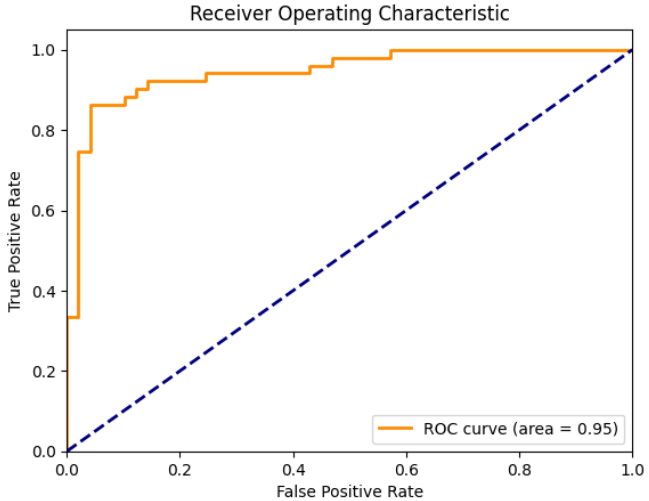

Figure 7. ROC Graph

The ROC image above shows the performance of the classification model with an AUC of 0.95. This shows that the model has a very good ability to distinguish between positive and negative classes. A fairly sharp curve approaching the upper left corner indicates that the model is able to achieve a high True Positive Rate with a low False Positive Rate, which is an indication of accurate predictions. However, the slight curvature of the curve indicates that although this model is very good, there are still some cases where the model experiences prediction errors. With an AUC of 0.95, this model is almost ideal, but still realistic in the context of real-world applications where some prediction errors are still possible. This shows a good balance between sensitivity and specificity in model predictions. Table 8 is a comparison of the accuracy of the machine learning algorithm with the proposed stacking model:

Table 8. Comparison of machine learning algorithm accuracy with stacking models

Model	Without SMOTE	SMOTE	ADABoost	Best Accuracy
SVM	82%	85%	88%	88%
LR	83%	82%	85%	85%
MLP	82%	84%	87%	87%
RF	84%	87%	89%	89%
Proposed Stacking Model	-	-	-	95%

Table 8 shows a comparison of the accuracy of several machine learning algorithms with the application of SMOTE, ADABoost, and the proposed stacking model techniques. From the table, it can be seen that the use of SMOTE successfully increased the accuracy for all models, with the largest increase occurring in the Random Forest model (from 84% to 87%). The application of ADABoost further increased the accuracy, especially in Random Forest which reached 89%. The proposed stacking model showed the best performance with the highest accuracy of 95%. This shows that the combination of various base models through the stacking technique is able to produce a more robust and accurate model than a single model that has been optimized with SMOTE, ADABoost, and XGBoost as a meta model. This confirms the superiority of the stacking approach in integrating the strengths of various algorithms to improve overall prediction accuracy. The following is table 9 which is a comparison of the accuracy of previous studies with the proposed model.

Table 9. Comparison with Previous Research

Researcher	Based Model	Meta Model	Accuracy
Rezaei Melal [24]	KNN, Decision Tree, RF, XGBoost	Neural Network (NN)	94.0%
Nyaramneni [25]	RF, XGBoost, LGBM	LR	94.7%
Almohimeed [26]	RF, DT, SVM, LR, KNN, NB	RF	90.03
Qian-Chuan [27]	KNN, RF, Support Vector Regression (SVR)	RF	93.8%
Kshatri [28]	SVM, J48, Naïve Bayes, Bagging, RF	SVM	94.5%
Seireg [29]	LGBM, GBR, XGBoost	Ridge Regression	93.5%
Our Model	SVM, LR, MLP, RF	XGBoost	95%

Table 9 presents a comparison of the results of the study with several previous studies that used various base models and meta models for classification. From the table, it can be seen that the model proposed in this study, which uses a combination of SVM, LR, MLP, and RF as base models and XGBoost as a meta model, achieves the highest accuracy of 95%. This is higher compared to previous studies involving various combinations of base models and meta models, with the previous highest accuracy of 94.7% achieved by Nvaramneni's study using RF, XGBoost, and LGBM as base models and Logistic Regression as a meta model [25]. These results indicate that the stacking approach proposed in this study is not only able to combine the strengths of multiple baseline models but also surpasses the performance of methods used in previous studies, thus offering a more accurate solution in the discussed classification context. Our proposed model outperforms previous studies due to its use of diverse base models (SVM, Logistic Regression, MLP, and Random Forest), capturing both linear and non-linear relationships in the data. XGBoost, chosen as the meta model, optimizes decision trees and corrects misclassifications, enhancing accuracy. By addressing class imbalance with SMOTE and incorporating ensemble techniques like K-Fold Cross Validation and boosting, our model achieves

superior performance, especially in handling imbalanced and complex physiological data, compared to earlier approaches.

While this research achieved a 95% accuracy in classifying student stress levels, several limitations could impact its performance and generalizability. The dataset, drawn from 2,851 university students in Riau Province, is geographically and demographically limited, potentially introducing bias. The physiological parameters used (SPO2, heart rate, temperature, blood pressure) may not fully capture stress variations across different populations or contexts. The model is highly specific to academic stress and may not generalize well to other stress environments. Despite using SMOTE to address class imbalance, synthetic data may not fully represent the complexity of stress states, leading to potential overfitting. The stacking ensemble model, while accurate, increases complexity and reduces interpretability, highlighting the need for explainability techniques. Additionally, the computational demands of the model may limit its scalability for larger datasets and broader applications.

The results of this research, which achieved 95% accuracy in classifying student stress levels, have significant practical implications for stress management at Hang Tuah University Pekanbaru and beyond. The model could enable early detection of stress, provide personalized interventions, and be integrated with wearable technology for continuous monitoring. It has broader applications in other educational settings, helping universities support at-risk students and improve mental health services. Additionally, the model could be adapted for non-academic environments, such as workplaces and healthcare settings, to monitor stress and prevent burnout.

To provide a clearer understanding of the practical aspects of the study, we detailed the implementation of the algorithms, including the software, hardware, and computational resources used. Python, along with libraries such as Scikit-learn, XGBoost, ADABoost, Pandas, and NumPy, was used for machine learning tasks, while Matplotlib and Seaborn were employed for visualizations. Development took place in Jupyter Notebook, and model training/testing was performed on an Intel® Core™ i9-11900H CPU with 16 GB of RAM, GPU (NVIDIA GeForce RTX 3060). Training the stacking ensemble model with 10-fold cross-validation took around 25-30 minutes, with an additional 5-10 minutes for SMOTE.

4. Conclusion

The conclusion of this research shows that the proposed stacking ensemble model successfully improves the accuracy in classifying and detecting stress levels of students at the Faculty of Computer Science, Hang Tuah University, Pekanbaru. By using a combination of basic models such as SVM, Logistic Regression, MLP, and Random Forest, and XGBoost as a meta-model, this model achieves an overall accuracy of 95%. The application of the SMOTE technique to overcome data imbalance increases the accuracy by up to 3% on some basic models, while the application of ADABoost successfully improves the accuracy by further 2-3% on each basic model. The application of K-Fold Cross Validation with K = 10 ensures that the resulting model has good generalization. In addition, the AUC test result of 0.95 shows that this model is very good at distinguishing between different stress classes, indicating high predictive ability. Overall, this stacking ensemble approach is proven to be superior with higher accuracy compared to traditional methods, as well as providing in-depth insights into the relationship between physiological indicators and stress levels of students. For further research, it is recommended to explore the use of deep learning-based meta-models such as LSTM (Long Short-Term Memory) and BiLSTM (Bidirectional LSTM). This would allow the model to capture subtle temporal variations in physiological signals that are important for distinguishing between stress levels such as Tense and Relaxed. These approaches have the potential to capture temporal patterns and long-term dependencies in physiological data that may not be fully optimized by traditional machine learning methods. In addition, future research can consider rotation techniques in stacking ensembles, where the base model algorithm can alternately act as a meta model. This approach can provide greater flexibility and allow for the exploration of more optimal model combinations, which can ultimately improve the overall prediction performance.

5. Declarations

5.1. Author Contributions

Conceptualization: H.F., Y.I., R.M., R.W., and A.M.; Methodology: Y.I. and R.M.; Software: H.F.; Validation: H.F., Y.I., R.M., R.W., and A.M.; Formal Analysis: H.F., Y.I., R.M., R.W., and A.M.; Investigation: H.F. and A.M.; Resources: Y.I. and R.W.; Data Curation: R.M. and R.W.; Writing Original Draft Preparation: H.F., Y.I., R.M., R.W., and A.M.; Writing Review and Editing: R.W. and A.M.; Visualization: H.F. and A.M.; All authors have read and agreed to the published version of the manuscript.

5.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

5.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

5.4. Institutional Review Board Statement

Not applicable.

5.5. Informed Consent Statement

Not applicable.

5.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Nemcova, "Monitoring of heart rate, blood oxygen saturation, and blood pressure using a smartphone," *Biomed. Signal Process. Control*, vol. 59, no. August, pp. 1–10, 2020, doi: 10.1016/j.bspc.2020.101928.

[2] Y. Irawan, Y. Fernando, and R. Wahyuni, "Detecting Heart Rate Using Pulse Sensor As Alternative Knowing Heart Condition," *J. Appl. Eng. Technol. Sci.*, vol. 1, no. 1, pp. 30–42, 2019, doi: 10.37385/jaets.v1i1.16.

[3] R. Wahyuni, Herianto, Ikhtiyaruddin, and Yuda Irawan, "IoT-Based Pulse Oximetry Design as Early Detection of Covid-19 Symptoms", *Int. J. Interact. Mob. Technol.*, vol. 17, no. 03, pp. 177–187, Feb. 2023.

[4] Herianto, B. Kurniawan, Z. H. Hartomi, Y. Irawan, and M. K. Anam, "Machine Learning Algorithm Optimization using Stacking Technique for Graduation Prediction," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1272–1285, 2024.

[5] T. Velmurugan and J. Dhinakaran, "A Novel Ensemble Stacking Learning Algorithm for Parkinson's Disease Prediction," *Math. Probl. Eng.*, vol. 2022, no. July, pp. 1-10, 2022, doi: 10.1155/2022/9209656.

[6] H. Zhang, H. A. Loaiciga, and T. Sauter, "A Novel Fusion-Based Methodology for Drought Forecasting," *Remote Sens.*, vol. 16, no. 5, pp. 1-25, 2024, doi: 10.3390/rs16050828.

[7] A. Ghasemieh, A. Lloyed, P. Bahrami, P. Vajar, and R. Kashef, "A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients," *Decis. Anal. J.*, vol. 7, no. May, pp. 1-13, 2023, doi: 10.1016/j.dajour.2023.100242.

[8] Z. Liao, M. Su, G. Ning, Y. Liu, T. Wang, and J. Zhou, "A Novel Stacked Generalization Ensemble-Based Hybrid PSVM-PMLP-MLR Model for Energy Consumption Prediction of Copper Foil Electrolytic Preparation," *IEEE Access*, vol. 9, no. January, pp. 5821–5831, 2021, doi: 10.1109/ACCESS.2020.3048714.

[9] M. Gollapalli et al., "A novel stacking ensemble for detecting three types of diabetes mellitus using a Saudi Arabian dataset: Pre-diabetes, T1DM, and T2DM," *Comput. Biol. Med.*, vol. 147, pp. 1-12, no. June, 2022, doi: 10.1016/j.compbiomed.2022.105757.

[10] S. Lin, X. Nong, J. Luo, and C. Wang, "A Novel Multi-Model Stacking Ensemble Learning Method for Metro Traction Energy Prediction," *IEEE Access*, vol. 10, no. November, pp. 129231–129244, 2022, doi: 10.1109/ACCESS.2022.3228441.

[11] S. K. Kalagotla, S. V. Gangashetty, and K. Giridhar, "A novel stacking technique for prediction of diabetes," *Comput. Biol. Med.*, vol. 135, no. February, pp. 1-11, 2021, doi: 10.1016/j.compbiomed.2021.104554.

[12] H. Zheng, S. W. A. Sherazi, and J. Y. Lee, "A Stacking Ensemble Prediction Model for the Occurrences of Major Adverse Cardiovascular Events in Patients with Acute Coronary Syndrome on Imbalanced Data," *IEEE Access*, vol. 9, no. July, pp. 113692–113704, 2021, doi: 10.1109/ACCESS.2021.3099795.

[13] C. Li, B. Xu, Z. Chen, X. Huang, J. He, and X. Xie, "A Stacking Model-Based Classification Algorithm Is Used to Predict Social Phobia," *Appl. Sci.*, vol. 14, no. 1, pp. 1-14, 2024, doi: 10.3390/app14010433.

[14] R. Forest, A. Boosting, and G. N. Bayes, "RAGN-L : A stacked ensemble learning technique for classification of Fire-Resistant columns," *Expert Systems with Applications*, vol. 240, no. September 2022, pp. 1-33, 2024, doi: 10.1016/j.eswa.2023.122491.

[15] M. A. Rahim, M. A. Hossain, M. N. Hossain, J. Shin, and K. S. Yun, "Stacked Ensemble-Based Type-2 Diabetes Prediction Using Machine Learning Techniques," *Ann. Emerg. Technol. Comput.*, vol. 7, no. 1, pp. 30–39, 2023, doi: 10.33166/AETiC.2023.01.003.

[16] H. Karamti, R. Alharthi, A. Al Anizi, R. M. Alhebshi, and A. A. Eshmawi, "Improving Prediction of Cervical Cancer Using KNN Imputed SMOTE Features and Multi-Model Ensemble Learning Approach," *Cancer*, vol. 15, no. 17, pp. 1–19, 2023.

[17] A. J. Mohammed, "Improving Classification Performance for a Novel Imbalanced Medical Dataset using SMOTE Method," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 9, no. 3, pp. 3161–3172, 2020, doi: 10.30534/ijatcse/2020/104932020.

[18] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on SMOTE algorithm," *Sci. Rep.*, vol. 11, no. 1, pp. 1–11, 2021, doi: 10.1038/s41598-021-03430-5.

[19] D. A. Kristiyanti, S. A. Sanjaya, V. C. Tjokro, and J. Suhali, "Dealing imbalance dataset problem in sentiment analysis of recession in Indonesia," *IAES Int. J. Artif. Intell.*, vol. 13, no. 2, pp. 2058–2070, 2024, doi: 10.11591/ijai.v13.i2.pp2060-2072.

[20] A. J. Barid, Hadiyanto, and A. Wibowo, "Optimization of the algorithms use ensemble and synthetic minority oversampling technique for air quality classification," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 33, no. 3, pp. 1632–1640, 2024, doi: 10.11591/ijeecs.v33.i3.pp1632-1640.

[21] T. R. Mahesh, V. K. V, D. K. V, O. Geman, and M. Margala, "Healthcare Analytics The stratified K-folds cross-validation and class-balancing methods with high-performance ensemble classifiers for breast cancer classification," *Healthc. Anal.*, vol. 4, no. July, pp. 1–10, 2023.

[22] C. Chen et al., "Improving protein-protein interactions prediction accuracy using XGBoost feature selection and stacked ensemble classifier," *Comput. Biol. Med.*, vol. 123, no. June, pp. 1-12, 2020, doi: 10.1016/j.compbiomed.2020.103899.

[23] M. Barton and B. Lennox, "Model stacking to improve prediction and variable importance robustness for soft sensor development," *Digit. Chem. Eng.*, vol. 3, no. May, pp. 1-13, 2022, doi: 10.1016/j.dche.2022.100034.

[24] S. Rezaei Melal, M. Aminian, and S. M. Shekarian, "A machine learning method based on stacking heterogeneous ensemble learning for prediction of indoor humidity of greenhouse," *J. Agric. Food Res.*, vol. 16, no. 1, pp. 1-12, 2024, doi: 10.1016/j.jafr.2024.101107.

[25] S. Nyaramneni, "ScienceDirect Advanced Ensemble Machine Learning Models to Predict SDN Advanced Ensemble Machine Learning Models to Predict SDN Tra ffi c Tra ffi," *Procedia Comput. Sci.*, vol. 230, no. December, pp. 417–426, 2024.

[26] A. Almohimeed et al., "Explainable Artificial Intelligence of Multi-Level Stacking Ensemble for Detection of Alzheimer's Disease Based on Particle Swarm Optimization and the Sub-Scores of Cognitive Biomarkers," *IEEE Access*, vol. 11, no. November, pp. 123173–123193, 2023, doi: 10.1109/ACCESS.2023.3328331.

[27] L. Qian-Chuan, X. Shi-Wei, Z. Jia-Yu, L. Jia-Jia, Z. Yi, and Z. Ze-Xi, "Ensemble learning prediction of soybean yields in China based on meteorological data," *Sci. Agric. Sin.*, vol. 22, no. 6, pp. 1909–1927, 2023, doi: 10.1016/j.jia.2023.02.011.

[28] S. S. Kshatri, D. Singh, B. Narain, S. Bhatia, M. T. Quasim, and G. R. Sinha, "An Empirical Analysis of Machine Learning Algorithms for Crime Prediction Using Stacked Generalization: An Ensemble Approach," *IEEE Access*, vol. 9, no. April, pp. 67488–67500, 2021, doi: 10.1109/ACCESS.2021.3075140.

[29] H. R. Seireg, Y. M. K. Omar, F. E. A. El-Samie, A. S. El-Fishawy, and A. Elmahalawy, "Ensemble Machine Learning

Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction," *IEEE Access*, vol. 10, no. June, pp. 64671–64687, 2022, doi: 10.1109/ACCESS.2022.3181970.

[30] M. K. Anam, S. Defit, Haviluddin, L. Efrizoni, and M. B. Firdaus, "Early Stopping on CNN-LSTM Development to Improve Classification Performance," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1175–1188, 2024, doi: 10.47738/jads.v5i3.312.

[31] W. Bourequat and H. Mourad, "Sentiment Analysis Approach for Analyzing iPhone Release using Support Vector Machine," *Int. J. Adv. Data Inf. Syst.*, vol. 2, no. 1, pp. 36–44, 2021, doi: 10.25008/ijadis.v2i1.1216.

[32] S. Mondal, R. Maity, Y. Omo, S. Ghosh, and A. Nag, "An Efficient Computational Risk Prediction Model of Heart Diseases Based on Dual-Stage Stacked Machine Learning Approaches," *IEEE Access*, vol. 12, no. January, pp. 7255–7270, 2024, doi: 10.1109/ACCESS.2024.3350996.

[33] A. Febriani, R. Wahyuni, Y. Irawan, and R. Melyanti, "Improved Hybrid Machine and Deep Learning Model for Optimization of Smart Egg Incubator," *J. Appl. Data Sci.*, vol. 5, no. 3, pp. 1052–1068, 2024.

[34] R. Perkasa, R. Wahyuni, R. Melyanti, Herianto, and Y. Irawan, "Light control using human body temperature based on arduino uno and PIR (Passive Infrared Receiver) sensor," *J. Robot. Control*, vol. 2, no. 4, pp. 307–310, 2021, doi: 10.18196/jrc.2497.