ISSN :2584-2706 DOI: https://doi.org/10.67317/IJMSRT26NOV056

Explainable Machine Learning for Multi-Class Mental Stress Level Prediction: A Comparative Analysis of Ensemble and Conventional Models

*Gourab Barui; Rajat Kumar Arya; Harjot Singh; Shivam Rana

Department of Computer Science and Engineering, ASET, Amity University, Noida, Uttar Pradesh, India

*Corresponding author

Abstract

Students' and professionals' mental stress has become one of the most serious public health issues; however, predictive models continue to be black boxes that do not get trusted by clinicians or teachers. In this research paper, we propose a comparison-based machine learning approach to solve the problem of classification of mental stress into three classes (Low, Moderate, High). Six machine learning algorithms such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, Random Forest, XGBoost, and LightGBM will be compared based on two feature spaces – all 20 features and the subset of 12 selected with the Mutual Information method. Evaluation was done based on the 1,100 samples survey dataset on psychological, physiological, social, environmental, and academic stressors. The best-performing single split was LightGBM, trained on features selected by the Mutual Information selection process, which resulted in an accuracy of 90.00%, Macro F1 score of 0.900, and Macro ROC-AUC of 0.987. It also showed the best mean cross-validation accuracy of 89.36% ± 0.84%. While tests show that the margin over Random Forest and Logistic Regression is directionally consistent but not statistically significant (as the number of samples is $n = 5$, therefore $p > 0.05$), it is statistically significant compared to XGBoost model (paired t-test, $p = 0.011$). To solve the problem of interpretability, SHapley Additive exPlanations (SHAP) were used in order to analyze the top-performing model. It showed that the key factors affecting students' stress level are blood pressure, self esteem, and anxiety level, followed by strong contributions from academic performance and depression. The results show that compact and interpretable ensembles can compete with and even surpass the performance of complex black-box models and provide a good basis for practical applications.

Keywords: *Predicting Mental Stress Levels, Explainable AI (SHAP), LightGBM, Features Selection, Machine Learning for Healthcare.*

I. INTRODUCTION

There has been a sharp rise in the case of mental stress among the student and professional community due to academic and occupational pressures, along with increased connectivity through digitization [1]. Chronic stress, which remains untreated and undetected, is highly likely to lead to other mental and physical problems like anxiety, depression, and cardiac issues. Machine learning techniques have recently been deployed for screening mental stress through tabular survey data, wherein the ensemble tree model has proven to be one of the best-performing models [3, 8].

Even though these models perform very well in terms of prediction, most stress classification algorithms based on machine learning remain as black-box models and do not provide any information regarding the reason why a risk has been predicted [12]. Explanatory Artificial Intelligence (XAI) techniques such as SHapley Additive exPlanation (SHAP) overcome this issue and have been largely used in recent clinical machine learning works [13, 14].

Contributions of this paper are: (i) benchmarking of six classical and ensemble classifiers for three-class (Low, Moderate, High) mental stress prediction using full-feature set and Mutual-Information selected feature set; (ii) determination of twelve features that perform equally well to the full-feature set and LightGBM trained on these twelve features gives the highest accuracy (90.00%) and macro F1-score (0.900) among all twenty four classifier-feature sets; and (iii) SHAP-based analysis of feature importance revealing that blood pressure, self esteem, and anxiety level are the key predictors of mental stress.

ISSN :2584-2706

II. RELATED WORK

2.1.Physiological and Wearable-Sensor Stress Detection

Stress detection through machine learning can be classified into physiological or wearable-sensor pipelines and survey-based or tabular pipelines. Early work by Gedam and Paul [3] and their subsequent review by Taskasap-lidis et al. [2], which further proves that ensemble or gradient-boosted classifier is a common classifier in recent literature, have covered the area of wearable sensor mental-stress detection. The scope of studies on personal-ized biosignals has been defined by Bolpagni et al. [4], while Chatzaki and Tsiknakis [5] have listed all available open physiological stress datasets and a systematic review from 2025 [6] reveals a spike in studies that examine stress using wearables and machine learning techniques since 2023.

2.2.Survey and Table-Based Academic Stress Prediction

Moving toward the more recent table-based approach, a study by Frontiers [7] compares the performance of five different classifiers on multi-datasets containing academic stress data, achieving a high accuracy range of 87.27-89.09%, along with an AUC-ROC range of 0.9853, in line with the accuracy achieved in this work. Several PMC studies have employed SHAP-based pipeline and improved decision trees on the same student stress dataset that is used in this work [8, 9], whereas another study on an associated corpus has achieved a higher accuracy of 93.48% through ensemble learning after oversampling [10] and while a benchmark test on 1,100 sample Kaggle stress corpus reveals that Random Forest is the best baseline (with up to 97.45%) [11].

2.3.Explanation-AI and Gradient-Boosted Ensembles

SHAP [15] continues to be the preeminent post-hoc model explainability approach in medical machine learning, used for workplace mental health predictions [12], depression risks derived from social media [13], and depression screening using NHANES data [14]. The two gradient boosting frameworks used in this study, XGBoost [16] and LightGBM [17], along with Random Forest [18], are considered state-of-the-art approaches to healthcare prediction and can therefore be used along with the standard baselines.

In light of the existing studies [8, 9], which have already implemented SHAP-based explanations for the 1,100 samples stress corpus, the significance of the current paper is not in the application of the XAI for this dataset for the first time but in the comparison of the six traditional and ensemble classifiers with full feature sets and with feature sets selected by mutual information, validated by nested cross-validation and paired statistical tests, and SHAP-based explanation of the best performing configuration.

III. PROPSED METHODOLOGY

3.1.Data Set and Data Preprocessing

The data set contains survey results of twenty psychological, physiological, social, environmental, and academic variables (for example, anxiety levels, self-esteem, blood pressure, quality of sleep, academic performance), and a three-level ordinal response variable (Low, Moderate, and High stress). The data were split 80/20 by using stratified sampling, and the features were normalized using the StandardScaler fitted only on the training data set.

3.2.Feature Selection based on Mutual Information

In order to examine whether a smaller, more understandable set of features is still predictive, the $K = 12$ most important features were selected based on MI (Mutual Information), which describes the dependence between feature X and label Y without linear correlation assumptions:

$$I(X; Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \left(\frac{p(x, y)}{p(x)\, p(y)} \right) \quad (1)$$

The selected twelve features constitute the Selected Features (K=12) configuration, benchmarked against the full All Features configuration.

3.3.Classification Models

Six classification algorithms including linear, distance, kernel, and ensemble techniques were fit on the same partitioning into training/testing subsets for each set of features. Logistic Regression learns the posterior prob-ability using multinomial softmax on standardized features (ℓ_2 penalty, $C = 10$). K-Nearest Neighbors (KNN) takes the majority of $k = 3$ closest neighbors weighted by their distance. Support Vector Machine (SVM) learns the separating boundary of maximal margin through an RBF kernel ($C = 10$). Random Forest learns 300 uncor-related trees from bootstrapped datasets and random feature subsets. XGBoost builds shallow trees (depth 8, learning rate 0.08) on the gradient of multinomial log-loss via second order boosting. LightGBM learns via histogram-based and leaf-wise gradient boosting (300 estimators, learning rate 0.08, depth 8).

ISSN :2584-2706 DOI: https://doi.org/10.67317/IJMSRT26NOV056

Table 1: Comparison of model performances for different feature sets (in order of Accuracy).

Model (Feature Set)	Acc.	Prec.	Rec.	F1	ROC-AUC
LightGBM (Selected K=12)	0.9000	0.9005	0.9004	0.9001	0.9871
Random Forest (All Features)	0.8955	0.8960	0.8958	0.8953	0.9849
Logistic Reg. (Selected K=12)	0.8909	0.8927	0.8914	0.8903	0.9833
XGBoost (Selected K=12)	0.8864	0.8865	0.8865	0.8865	0.9843
SVM (Selected K=12)	0.8818	0.8822	0.8824	0.8820	0.9736
Random Forest (Selected K=12)	0.8818	0.8824	0.8821	0.8815	0.9837
Logistic Regression (All Features)	0.8818	0.8819	0.8821	0.8818	0.9853
LightGBM (All Features)	0.8773	0.8775	0.8776	0.8768	0.9836
KNN (Selected K=12)	0.8727	0.8765	0.8734	0.8725	0.9531
SVM (All Features)	0.8727	0.8748	0.8732	0.8718	0.9839
XGBoost (All Features)	0.8682	0.8682	0.8684	0.8683	0.9826
KNN (All Features)	0.8545	0.8569	0.8551	0.8546	0.9363

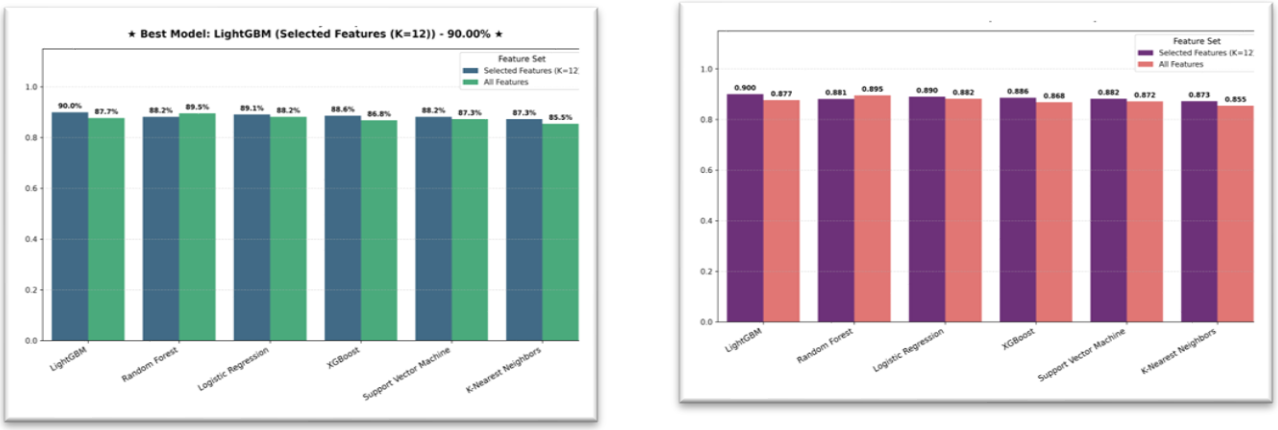


Fig. 1: Accuracy (left) and macro F1-score (right) comparison across all six classifiers under both feature con-figurations. LightGBM with Mutual-Information-selected features (K=12) achieves the top accuracy of 90.00%.

3.4.Cross-Validation Protocol and Mitigation of Data Leakage

To check whether the result of the 80:20 split held-out approach is an artifact of good partitioning, all twelve models and feature sets were tested in stratified 5-fold cross-validation (n = 5, random state=42). For each fold, standardization and Mutual-Information based feature selection was fitted on the training subset only and then used on the held-out subset – that is, feature selection takes place within the cross-validation loop instead of fitting on the whole data – so that no information from the test folds would influence model fitting and feature selection. The accuracy, macro F1 score, and macro ROC-AUC were measured for each combination of models and configuration and the best performing ones were compared with a paired t-test and the non-parametric Wilcoxon signed-rank test on the fold accuracy

.

3.5.Interpretability of the Explanatory Model using the SHAP Framework

In order to interpret the best-performing model, the SHAP value [15] technique was used. SHAP assigns a contribution value ϕ_i to each feature i, which corresponds to the average contribution to the model output by that particular feature for all subsets S of features $N \setminus \{i\}$, according to cooperative game theory:

$$\phi_i(v) = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!\,(|N| - |S| - 1)!}{|N|!} \left(v(S \cup \{i\}) - v(S) \right) \quad (2)$$

where N is the complete feature set, while $v(S)$ is the predicted value of the model based on the conditional features S. With respect to the winning tree-based model, SHAP was estimated using the TreeExplainer, providing local attributions as well as global SHAP rankings for all classes.

IV. RESULTS

4.1.Overall Model Performance Comparison

The Accuracy, macro Precision, macro Recall, macro F1-Score, and macro ROC-AUC scores for all six models for both sets of features are presented in Table 1. LightGBM using Mutual-Information selected twelve features was the best performing model (Accuracy = 90.00%, macro F1-Score = 0.900, macro ROC-AUC = 0.987). The next-best performing model was Random Forest with complete set of features (Accuracy = 89.55%), followed by Logistic Regression using Mutual-Information selected set of twelve features (Accuracy = 89.09%). All models had high ROC-AUC values in excess of 0.97 for their macro performance metrics, which signifies good class separability regardless of architecture.

Fig. 1 shows accuracy and macro F1-score performance for both feature sets. Feature selection helped increase accuracy and F1-score of four out of six models (LightGBM: +2.3 pp, Logistic Regression: +0.9 pp, XGBoost:

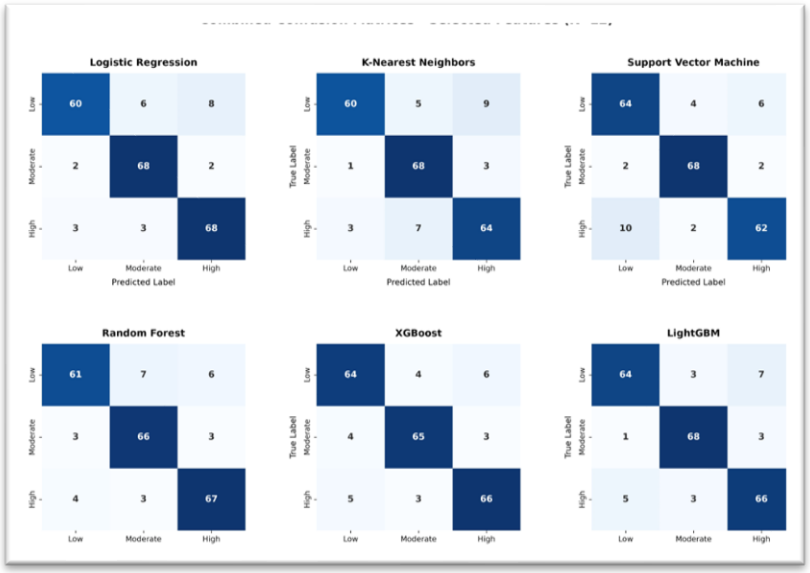


Fig. 2: Combined confusion matrices for all six classifiers, Selected Features (K=12).

Table 2: Accuracy of 5-fold stratified cross-validation ($\mu \pm \sigma$, $n = 5$) for the best models and features combinations, nested MI selection in each fold.

Model (Feature Set)	CV Accuracy (%)
LightGBM (Selected K=12)	89.36 ± 0.84
Logistic Regression (All)	89.00 ± 2.16
XGBoost (Selected K=12)	88.45 ± 0.94
Random Forest (Selected K=12)	88.18 ± 0.76
Logistic Regression (K=12)	88.00 ± 1.83
Random Forest (All)	87.82 ± 1.50

+1.8 pp, KNN: +1.8 pp) and the use of all features only helped Random Forest slightly, probably due to the built-in bootstrapping and subsampling.

4.2.Class-wise Error Analysis

Fig. 2 displays combined confusion matrices for six models (Selected Features, K=12). The best diagonal is provided by LightGBM (64/68/66 correct in Low/Moderate/High) with remaining confusions distributed around Low and High ends of stress-level spectrum but not the adjacent classes. Most stable classification is achieved for the Moderate class across all models. SVM shows the highest asymmetric confusions, confusing 10 High-stress observations with Low category, suggesting that RBF classifier boundary is less sensitive to feature distribution extrema than tree-based ensembles.

ISSN :2584-2706 DOI: https://doi.org/10.67317/IJMSRT26NOV056

4.3.Cross-Validation and Statistical Significance

The results for 5-fold cross-validated accuracy means ± standard deviations for the best model-feature com-binations can be found in Table 2 based on the nested (leakage-free) method described in Section 3.4. In the cross-validation setting, the LightGBM (Selected K=12) model once again achieves the maximum mean accu-racy (89.36% ± 0.84%), followed by Logistic Regression (All Features, 89.00% ± 2.16%) and XGBoost (Selected K=12, 88.45% ± 0.94%); the complete ranking of all twelve configurations matches but does not match the single split ranking in Table 1. This verifies that the single split ranking was not a favorable choice of partition. Pairwise significance tests of fold-wise accuracy suggest that LightGBM (K=12) significantly outperforms Random Forest (All) by 1.55 percentage points (paired t-test $p = 0.082$, Wilcoxon $p = 0.125$), and significantly outperforms Logistic Regression (K=12) by 1.36 percentage points (both $p = 0.125$ and $p = 0.188$). However, the direction of the differences is positive and these three models show comparable performance on the given scale. Meanwhile, 0.91 percentage point difference between LightGBM (K=12) and XGBoost (K=12) is sta-tistically significant in the paired t-test ($p = 0.011$) but not in the Wilcoxon test ($p = 0.063$), suggesting low statistical power of a 5-fold test. Thus, we consider LightGBM (K=12) as the model with the best accuracy and macro F1 score, which is indistinguishable from Random Forest and Logistic Regression under the top feature configurations and recommend testing it with a larger validation sample (Section 6).

V. EXPLANATIONS THROUGH SHAP

The global SHAP feature impact ranking of the winning LightGBM classifier, split into classes (0=Low, 1=Moderate, 2=High), is displayed in Fig. 3, (left). Blood pressure is the only feature of strong influence, followed by self esteem and anxiety level; academic performance and depression make up a second layer of influence, while social support and extracurricular activities have the weakest contributions out of all twelve

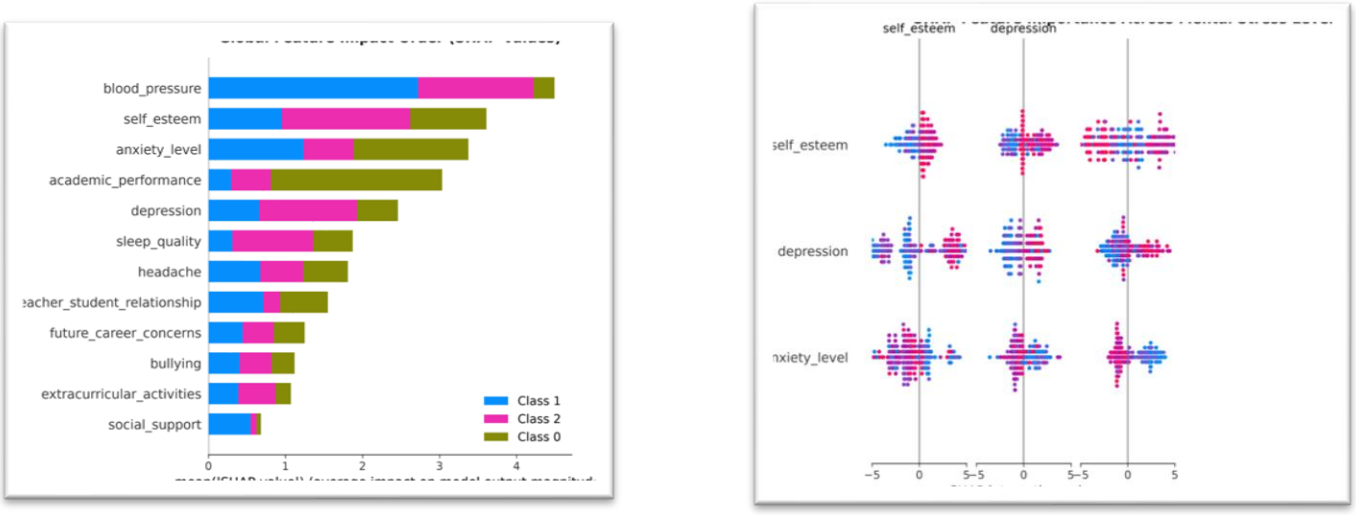


Fig. 3: Left: global SHAP feature ranking (mean |SHAP value|) by class, LightGBM (K=12). Right: SHAP interaction values among top features (self esteem, depression, anxiety level).

chosen features. This ranking is consistent with clinical knowledge: high blood pressure is a proven indicator of sympathetic-nervous-system activity under chronic stress, whereas low self esteem and high anxiety level are primary components, not simply secondary indicators. The importance of academic performance is a con-sequence of the academic nature of the sample, and the contribution of depression suggests its presence as a feature-level comorbidity.

The right plot in Fig. 3 displays SHAP interaction-style plots for the top three features colored according to predicted labels. Features with the broadest interaction range against depression are self esteem and anxi-ety level, meaning that the model works with the combination of psychological features rather than individual feature value, which is the strength of tree-based models over logistic regression.

VI. CONCLUSION AND FUTURE

This research compared the performances of six ML algorithms for the multi-class classification of mental stress in the presence of both full and Mutual-Information selected features, and found that the use of LightGBM on the reduced set of twelve features results in the highest split-accuracy (90.00%) and the highest average of 5-fold cross-validation accuracy (89.36% ± 0.84%). A pairwise test for statistical significance revealed that this configuration is not significantly different from those using Random Forest (All Features) and Logistic Regression ($p > 0.05$), hence we designate LightGBM (K=12) to be the most effective configuration by mean accuracy but not the superior one – a

crucial consideration for such a limited-size dataset. Using SHAP for model interpretation, the study found out that physiological (blood pressure) and psychological (self esteem, anxiety level) parameters are the most important for model predictions and thus provide clinical interpretation and justification for stress screening applications.

Future work will (i) extend validation of the proposed framework in large, external, and diverse groups to enhance statistical power beyond current $n = 5$-fold methodology, (ii) integrate time-series and sensor-based data along with tabular features to track the dynamics of stress, and (iii) investigate probability estimates and fairness-aware evaluation before deployment in the real world.

REFERENCES

[1]. A.-F.A. Mentis, D. Lee, and P. Roussos, "Applications of artificial intelligence-machine learning for detection of stress: a critical overview," Molecular Psychiatry, vol. 29, no. 6, pp. 1882–1894, 2024.
[2]. G. Taskasaplidis, D.A. Fotiadis, and P.D. Bamidis, "Review of stress detection methods using wearable sensors," IEEE Access, vol. 12, pp. 38219–38246, 2024.
[3]. S. Gedam and S. Paul, "A review on mental stress detection using wearable sensors and machine learning techniques," IEEE Access, vol. 9, pp. 84045–84066, 2021.
[4]. M. Bolpagni, S. Pardini, M. Dianti, and S. Gabrielli, "Personalized stress detection using biosignals from wearables: a scoping review," Sensors, vol. 24, no. 10, p. 3221, 2024.
[5]. C. Chatzaki and M. Tsiknakis, "An overview of stress analysis based on physiological signals: systematic review of open datasets and current trends," Sensors, vol. 25, p. 7108, 2025.
[6]. "Use of machine learning for predicting stress episodes based on wearable sensor data: a systematic review," Computers in Biology and Medicine, vol. 198, Part A, 2025.
[7]. "An explainable machine learning framework for academic stress classification among university students: a compar-ative multi-dataset study with ablation analysis and statistical significance testing," Frontiers in Computer Science, vol. 8, 2026.
[8]. "Explainable artificial intelligence for predictive modeling of student stress in higher education," PMC, National Library of Medicine, 2026.
[9]. "Students' stress prediction and explainable analysis based on improved decision trees," PMC, National Library of Medicine, 2026.
[10]. "Enhancing diagnostic decision-making: ensemble learning techniques for reliable stress level classification," Diagnostics, vol. 13, no. 22, p. 3455, 2023.
[11]. "Predictive modeling of students' stress levels using machine learning algorithm," ResearchGate preprint, 2025.
[12]. "Leveraging machine learning algorithms and explainable AI for predicting mental health disorder treatment at the workplace," ScienceDirect, 2026.
[13]. "Explainable machine learning for mental health prediction from social media behavior: a nested cross-validation study with SHAP and LIME interpretability," Discover Mental Health, Springer Nature, 2026.
[14]. "Development of an explainable machine learning model for predicting depression in adults with type 2 diabetes mellitus: a cross-sectional SHAP-based analysis of NHANES 2009–2023," Medicine, vol. 105, 2026.
[15]. S.M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.
[16]. T. Chen and C. Guestrin, "XGBoost: a scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, pp. 785–794, 2016.
[17]. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: a highly efficient gradient boosting decision tree," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 3146–3154, 2017.
[18]. L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.