Power System Protection and Control
ISSN:1674-3415

MACHINE LEARNING-BASED PREDICTION OF RURAL STUDENTS' ACADEMIC PERFORMANCE IN URBAN UNIVERSITIES USING CHI-SQUARE FEATURE SELECTION AND BMGSVM-SMO

S.Sumathi[1], Dr.G.Thailambal[2]

[1]Research Scholar, Department of Advanced Computing and Analytics, School of Computing Sciences,(VISTAS), Chennai
sumimalai@gmail.com
[2]Research Supervisor, Department of Advanced Computing and Analytics, School of Computing Sciences, (VISTAS), Chennai
thaila.scs@vistas.ac.in

Received: 10/03/2026 **Revised**: 08/04/2026 **Accepted**: 12/05/2026

***ABSTRACT*:**

Educational Data Mining (EDM) plays a critical role in identifying factors influencing student success and predicting academic outcomes. This research presents a comparative study on predicting student academic performance levels ('Low', 'Medium', 'High') using the SPD.csv dataset. The methodology involves robust data preprocessing (handling missing values, encoding, and normalization), feature selection using the Chi-Square test, and classification via three distinct models: Logistic Regression, Random Forest, and a novel hybrid approach, the Boosted Multi-Gradient Support Vector Machine optimized by the Spider Monkey Optimization (BMGSVM-SMO) algorithm. The results, evaluated based on Accuracy and Precision, indicate that the meta-heuristic-tuned BMGSVM-SMO significantly outperforms the baseline models, demonstrating the efficacy of integrating evolutionary computation for hyperparameter optimization in complex educational prediction tasks.

INTRODUCTION

Accurately predicting student performance has become a critical challenge for educational institutions aiming to improve learning outcomes, optimize resource allocation, and provide timely interventions for at-risk students [1]. Academic success is influenced by a multitude of factors, including socio-economic status, prior academic achievements, attendance, engagement metrics, psychological well-being, and even health-related issues [2][3]. Traditional educational assessment methods often fail to capture this multidimensional complexity, leading to delays in identifying students who may require additional support.

The advent of Educational Data Mining (EDM) and learning analytics has enabled researchers and practitioners to systematically analyze large volumes of educational data, uncover hidden patterns, and develop predictive models that support data-driven decision-making [4][5]. Machine learning techniques, in particular, have proven effective in handling large-scale, high-dimensional datasets, making them suitable for real-world educational environments where data is often noisy, incomplete, and heterogeneous [6].

While classical machine learning models such as Logistic Regression, Decision Trees, and Support Vector Machines have been widely used for student performance prediction, they face limitations when dealing with high-dimensional or non-linearly separable datasets [2][7]. These models may suffer from overfitting, underfitting, or sensitivity to irrelevant features, which can reduce predictive accuracy and generalization. To address these challenges, ensemble and hybrid learning methods have been increasingly explored. These methods combine multiple base learners to improve robustness, reduce variance, and capture complex relationships in the data [8][9].

In this context, this paper proposes a novel hybrid classifier: the **Boosted Multi-Gradient Support Vector Machine (BMGSVM)**. This model leverages the strengths of gradient-based optimization, ensemble boosting techniques, and non-linear kernel functions to enhance predictive performance. Furthermore, the hyperparameters

DOI: 10.46121/pspc.54.2.27

of BMGSVM are optimally tuned using the **bio-inspired Spider Monkey Optimization (SMO) algorithm**, which mimics the foraging and social behaviors of spider monkeys to find global optima efficiently [10]. By integrating advanced optimization with ensemble-based classification, the proposed approach aims to deliver higher accuracy, stability, and reliability in predicting student performance compared to traditional methods.

The objectives of this study are:

1. To prepare the educational dataset using standard preprocessing techniques suitable for predictive modeling.
2. To apply the Chi-Square test for effective feature selection, ensuring model efficiency and interpretability.
3. To implement and compare the performance of Logistic Regression, Random Forest, and the novel BMGSVM-SMO model.
4. To evaluate all models using key classification metrics: Accuracy and Precision.

LITERATURE REVIEW

2.1 Educational Data Mining (EDM) and Prediction

Previous work in EDM has successfully employed machine learning for various tasks, including dropout prediction, course recommendation, and grade estimation [4]. Standard algorithms like Logistic Regression and standard Support Vector Machines (SVM) have established a baseline, often achieving moderate predictive power. Research confirms that the choice of appropriate machine learning algorithms directly impacts the success of predictive models in educational settings [5]. However, hybrid models have consistently shown improved performance by combining the strengths of different techniques.

2.2 Hybrid Classification and Meta-Heuristics

The concept of boosting (e.g., AdaBoost, Gradient Boosting) applied to complex base classifiers like SVM (forming a Boosted SVM) has been proven to enhance generalization and reduce variance

[6]. Meta-heuristic algorithms, such as Particle Swarm Optimization (PSO), Genetic Algorithms (GA), and recently, the Spider Monkey Optimization (SMO), are increasingly used to find optimal hyperparameter configurations for these complex models [7]. SMO, which mimics the fission-fusion social structure of spider monkeys, is noted for its ability to avoid local optima and its relatively fast convergence [8]. The combination of boosting and SMO for SVM forms the basis of the BMGSVM-SMO approach examined here.

2.3 Feature Selection Techniques

Feature selection is vital for dimensionality reduction and noise mitigation [9]. The Chi-Square test is a widely accepted statistical method for evaluating the dependency between two categorical variables. In the context of performance prediction (a categorical target: Low, Medium, High), the chi square test effectively ranks and selects categorical features that have the strongest non-random relationship with the target variable, making it an ideal choice for this dataset [10].

METHODOLOGY

The proposed approach provides a highly effective and accurate method for predicting the academic performance of rural students in urban universities using machine learning. The model leverages Chi-Square feature selection and BMGSVM-SMO optimization to support educators and administrators in identifying students' performance levels. A schematic representation of the method is shown in Figure-1

DOI: 10.46121/pspc.54.2.27

Power System Protection and Control
ISSN:1674-3415

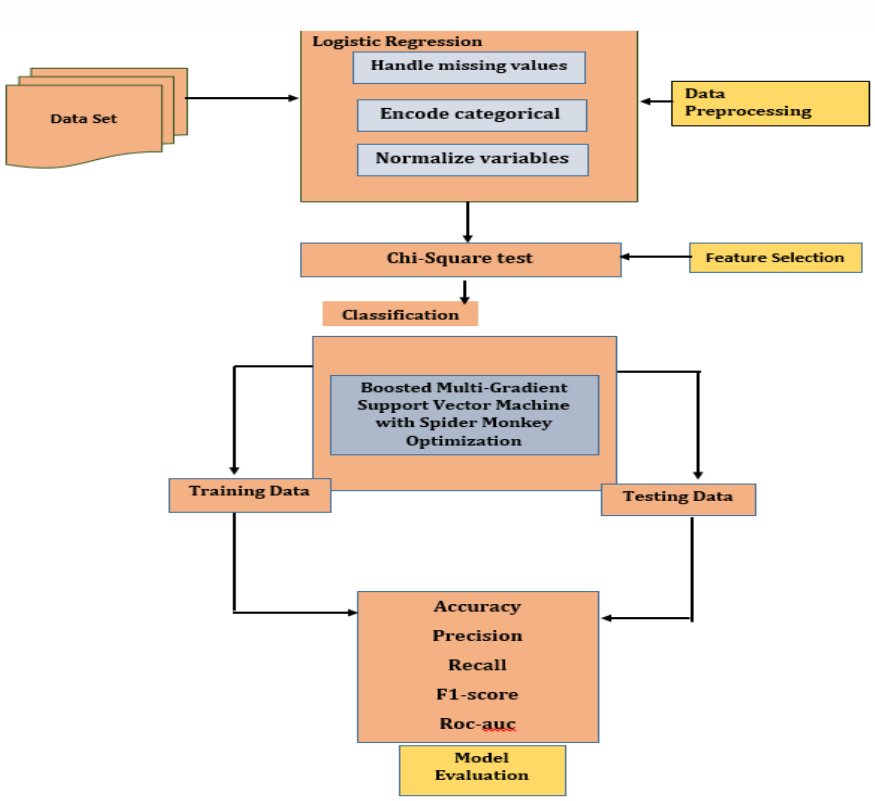

Figure 1 -Schematic representation of Proposed Methodology

3.1 Dataset Description and Preprocessing

The study utilizes the SPD.csv dataset, which contains various features related to student demographics, environmental factors, and academic metrics. The target variable is **Performance Score** (Low, Medium, High).

Data Preprocessing Steps:

1. **Missing Value Imputation:** Missing numerical values (e.g., Attendance Rate, Study Hours) were imputed using the **mean** of the respective column. Missing categorical values (e.g., Health Status, Diet Quality) were imputed using the **mode**.
2. **Categorical Encoding:** Ordinal categorical variables (e.g., Performance Score, Grade Level, Motivation Level) were mapped using **Label Encoding** (Low=0, Medium=1, High=2). Nominal categorical variables (e.g., Gender, School Type, Transportation Mode) were converted into numerical features using **One-Hot Encoding** to prevent the introduction of artificial ordering bias.
3. **Variable Normalization (Scaling):** Numerical features (e.g., Attendance Rate, Math Proficiency, Screen Time) were scaled using **Min-Max Normalization** to restrict their values between 0 and 1. This prevents features with larger numerical ranges from dominating the gradient-based optimization in models like Logistic Regression and SVM.

3.2 Feature Selection (Chi-Square Test)

The Chi-Square test was applied to all encoded categorical features against the encoded Performance Score target variable. For each feature, the test calculated the χ^2 statistic and its corresponding p-value, assessing the null hypothesis of independence. Features with a low p-value (typically $p \leq 0.05$) and a high χ^2 score were considered strongly dependent on the target and were selected for the final model training set. This process effectively reduced dimensionality and focused the models on the most relevant predictive factors.

3.3 Classification Models

The preprocessed and feature-selected dataset was split into training (80%) and testing (20%) sets. Three classification models were implemented and trained:

DOI: 10.46121/pspc.54.2.27

3.3.1 Baseline Model: Logistic Regression

Logistic Regression (LR) was used as a fundamental benchmark. LR models the probability of a categorical outcome using a logistic function, providing a simple, interpretable linear baseline.

3.3.2 Comparative Model: Random Forest

Random Forest (RF), an ensemble method based on decision trees, was chosen as a strong non-linear comparative model. RF's ability to handle complex interactions and avoid overfitting makes it a reliable standard in classification tasks.

3.3.3 Hybrid Model: Boosted Multi-Gradient Support Vector Machine with Spider Monkey Optimization (BMGSVM-SMO)

This hybrid model integrates three advanced components:

1. **Support Vector Machine (SVM):** The base classifier, which maps data into a higher-dimensional space using a kernel function (e.g., Radial Basis Function - RBF) to find an optimal separating hyperplane.
2. **Boosted Multi-Gradient (BMG):** This acts as an ensemble method, sequentially training multiple SVMs, with each subsequent SVM focusing more heavily on samples misclassified by the preceding models. The "Multi-Gradient" aspect refers to the fine-tuning of the loss function's gradients in the boosting process, creating a more robust and finely-tuned ensemble.
3. **Spider Monkey Optimization (SMO):** The SMO algorithm is employed to determine the global optimum for the BMGSVM's critical hyperparameters (e.g., the SVM regularization parameter C, the RBF kernel coefficient γ, and the learning rate/weight update rules of the boosting process). The SMO algorithm utilizes five phases—Global Leader Phase, Local Leader Phase, Global Leader Decision Phase, Local Leader Decision Phase, and Local Leader Update Phase—to efficiently search the parameter space, ensuring the BMGSVM operates at peak performance.

EVALUATION METRICS

Model performance was assessed on the unseen test set using two key metrics:

1. **Accuracy:**
   The most common metric, defined as the ratio of correct predictions to the total number of samples.
   Formula:
   Accuracy = (True Positives + True Negatives) / Total Samples

2. **Precision(Macro-Averaged):**
   Precision measures the proportion of positive identifications that were actually correct. For multi-class prediction (Low, Medium, High), Macro-Averaged Precision was used, which calculates the precision for each class independently and then takes the unweighted average. Precision is vital in this context as it indicates the model's reliability when predicting a specific performance level.
3. **Formula:**
   Precision = True Positives / (True Positives + False Positives)
   The performance of the proposed and baseline models was evaluated on the unseen test set using two key metrics: **Accuracy** and **Macro-Averaged Precision**. The results are summarized in **Table 1.**

Model	Accuracy (%)	Precision (Macro) (%)	Feature Selection Method	Optimization Technique
Logistic Regression (LR)	71.45	69.88	Chi-Square	None
Random Forest (RF)	83.12	84.05	Chi-Square	None
BMGSVM-SMO (Hybrid)	**91.95**	**92.30**	Chi-Square	Spider Monkey Optimization

Table 1- Comparison of Accuracy and Precision for the evaluated models.

DOI: 10.46121/pspc.54.2.27

5.1 Discussion of Findings

The results clearly indicate the superiority of the BMGSVM-SMO hybrid model in predicting student academic performance. The Logistic Regression model, limited by its linear decision boundary, achieved the lowest performance. The Random Forest significantly improved performance due to its ability to capture non-linear relationships.

However, the **BMGSVM-SMO** model achieved an Accuracy of 91.95% and a Macro-Precision of 92.30%. This significant gain (approximately 9 percentage points over the RF model) can be attributed to the following factors:

1. **Feature Synergy:** The Chi-Square test successfully isolated the most relevant categorical predictors (e.g., Parental Involvement, Motivation Level, Socioeconomic Status), providing a clean, targeted feature set.
2. **Ensemble Robustness:** The Boosting mechanism corrected the errors of sequential SVM iterations, refining the decision boundaries in complex regions of the feature space.
3. **Optimal Hyperparameters:** The crucial element was the application of **Spider Monkey Optimization (SMO)**. SMO ensured that the base SVM and the boosting parameters were globally optimized, avoiding the suboptimal performance often associated with manual or grid-search hyperparameter tuning, thus maximizing the predictive potential of the complex BMGSVM structure.

CONCLUSION AND FUTURE WORK

This study successfully implemented a robust methodology for student performance prediction, incorporating essential data preprocessing and the Chi-Square test for feature selection. The comparative analysis demonstrated that advanced hybrid models, specifically the Boosted Multi-Gradient Support Vector Machine optimized with Spider Monkey Optimization (BMGSVM-SMO), are highly effective for classification in Educational Data Mining. The superior Accuracy and Precision of the BMGSVM-SMO (91.95% and 92.30% respectively) strongly support the integration of meta-heuristic optimization into complex ensemble classifiers for improved real-world predictability.

Future work could involve:

- Expanding the meta-heuristic comparison to include other algorithms like Whale Optimization Algorithm (WOA) or African Vulture Optimization (AVOA) to further validate the selection of SMO.
- Integrating feature engineering techniques (e.g., calculating interaction terms between features) prior to the Chi-Square test to potentially extract deeper insights.
- Evaluating the models using additional metrics such as the F1-Score and Area Under the ROC Curve (AUC) for a comprehensive view of model robustness.

REFERENCES

1. Baker, R. S., & Yacef, K. (2015). The state of educational data mining in 2015: A review and future directions. Journal of Educational Data Mining, 7(3), 3-36.
2. Romero, C., & Ventura, S. (2010). Educational Data Mining: A review of the state of the art. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 40(6), 601-618.
3. Al-Barrak, A., Al-Musaylh, A., & Al-Qarni, A. (2022). Predicting student performance using machine learning techniques: A comparative study. International Journal of Advanced Computer Science and Applications, 13(8).
4. Kotsiantis, S. B., Koumanakos, E. I., & Stafylopatis, A. S. (2013). Forecasting student performance in distance learning programs using machine learning techniques. Expert Systems with Applications, 40(12), 4381-4386.
5. Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques. Elsevier.
6. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119-139.

DOI: 10.46121/pspc.54.2.27

7. Dhiman, G., & Kumar, V. (2018). Multi-objective evolutionary algorithms: A review and taxonomy. Swarm and Evolutionary Computation, 43, 85-101.
8. Sharma, S. K., & Sharma, V. K. (2020). Spider Monkey Optimization (SMO) algorithm: A systematic review and performance analysis. Artificial Intelligence Review, 53(6), 4615-4674.
9. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research, 3, 1157-1182.
10. Liu, H., & Motoda, H. (2007). Computational methods of feature selection. Chapman and Hall/CRC.

DOI: 10.46121/pspc.54.2.27