iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** XI **Month of publication:** November 2022

DOI: https://doi.org/10.22214/ijraset.2022.47485

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue XI Nov 2022- Available at www.ijraset.com

Enticing Correlation between Motivation and Self-Regulation: A Novel Approach for Academic Excellence

Latika Kharb[1], Deepak Chahal[2]
[1, 2]*Jagan Institute of Management Studies, Sector-5, Rohini, Delhi-110085, India*

Abstract: ***Today, we can easily say that there is no limit to the size of the data that exists in the real world-it can be in tens, hundreds, thousands, billions or even trillions of dimensions. However, not all dimensions are vital or relevant for a more detailed analysis. The key objective of educational data mining is to analyze educational data to solve educational and research problems. Our research in this paper focuses primarily on identifying relevant characteristics that can help determine the academic performance of students. The selection of features carried out as data preprocessing techniques keeping in mind the key idea to eliminate irrelevant and redundant features and thus selecting the optimal characteristics that can improve the overall accuracy of the model. The characteristics selected through the selection of functions can help to predict the academic performance of the student in several colleges and universities, which is one of the most imperative contemplations by interested parties to build and maintain a quality system. Although there are several studies that determine the factors that affect the academic performance of students, there is still a gap in existing research that focuses on the psychological factors of a student. The document is structured as follows. Section two contains a brief explanation about the related work in this field. In section three, we have discussed the adopted methodology that presents the data set, the feature selection algorithms, and the characteristic evaluation approach. Section four presents the configuration and experimental results and section five focuses on the conclusion and future scope.***

Keywords: Personality traits, feature selection, algorithm, classifier, academic performance.

I. INTRODUCTION

In order to keep track of the changes occurring in curriculum patterns, a regular analysis is a must in the educational field [1]. To determine the main influencing characteristics that affect the academic accomplishment of the students, we have contemplated the previous efforts of other researchers in the field. Our research mainly considers the psychological factors that influence the student's academic performance. Instructors and others seriously need to have relevant and timely information about higher Education's overall performance which is all about Students' academic activity in higher educations [2]. Researchers have recently proposed several machine learning-based algorithms for predicting academic achievement [3]. Michelle Richardson et al. conducted a 13-year review, which focused on student performance based on the Grade Point Average (GPA). He studied three important characteristics, psychological and demographic, that affect the student's academic performance. In their research, they have identified awareness, the need for cognition, emotional intelligence, the place of control, optimism, intrinsic academic motivation, orientation of learning objectives, regulation of effort, anxiety about exams, measures of commitment of objectives, general stress and academic stress. GPA Among the demographic features, it was observed that the older and higher socioeconomic students obtained higher grades.

In addition, Narges Babakhani et al. studied the relationship between personality traits (kindness, extraversion, openness, neuroticism and awareness), learning strategies and academic performance. The result confirmed that all personality factors (except neuroticism) and self-regulated learning strategies are significant predictors of academic achievement. RajuRanjan et al. developed an effective model to predict the effect of several parameters identified in the CGPA of students using Logistic Regression. Thirteen parameters were used to construct the model. The result showed that age was the strongest predictor of CGPA followed by the student's educational history, closeness to the family and the freedom to make decisions and the family's economic environment. MeeraKomarraju et al. conducted research to identify the association between personality traits, academic performance and motivation among college students in the United States. It was concluded in his study that students with high consciousness, openness, kindness, strong aspiration and less neurotic are expected to have a higher GPA. Identifying the most important attributes that affect the performance of students is one of the main objectives of the research conducted by Amirah Mohamed Shahiria et al.

Internal assessment, demographic factors, external evaluation, psychometric factors, extracurricular activities, interaction of social networks, social skills are the attributes that were studied and modeled using various data mining techniques. Data Science involves developing techniques of storing, recording and analyzing data to extract useful information efficiently [4]. It was observed that the neural network had the highest prediction accuracy. T. Bidjerano et al. in their study identified the correlation between the personality traits of the five and the use of self-regulated learning strategies. His research revealed that there is an independent contribution of the personality trait of the intellect and the student's grade point average, while the regulation of effort arbitrated the effects of kindness and awareness. The decision tree approach is used by K.ShanmugaPriya et al. in his study to improve student performance. They concluded that the student's academic performance in an educational system can be improved by analyzing communication skills, paper presentations, and internal and final semester evaluation. Naïve Bayes, Decision Trees and Neural Network were applied by E. Osmanbegović et al. in their research on preoperative assessment data to predict the achievement of the course. The performance of the learning methods was evaluated on the basis of predictive accuracy, user-friendly features and ease of learning. The research study conducted by Margarete Imhof et al. addressed the importance of motivational and cognitive aspects and their impact on student learning outcomes. Their results showed that prior knowledge, situational interest, self-concept, learning strategies and awareness based on effort and effort; All are positively related to the final performance. Table 1 below describes the most studied Influence Factors in previous research that affect the student's academic performance.

II. METHODOLOGY AND APPROACH

The main objective of our research is to determine the most significant and optimal characteristics that can help to predict the academic performance of students at the graduate and postgraduate levels. An empirical study of the related area was conducted to find the existing characteristics / parameters and how well they help to predict the CGPA (cumulative grade point average). Educational data mining has become an effective tool for exploring the hidden relationships in educational data and predicting students' academic achievements [5]. The preprocessing of data, the preliminary and one of the most important steps in data mining includes cleaning, integration, transformation, feature extraction and selection. The selection of features, which is one of the main aspects of data preprocessing and an imperative part of machine learning, is gaining popularity, as it helps to identify the most relevant characteristics, which reduces excessive adjustment, reduces training time and improves the prediction performance. Feature selection is the process of identifying the most relevant features from the given data having a large feature space [6] The selection of features not only helps improve the quality of the model, but also makes the modeling process more efficient. Feature selection (FS) seeks to enhance classification efficiency by selecting only a tiny subset of appropriate features from the initial wide range of features [7]. Depending on the selection algorithm and the creation of models, the selection of characteristics is generally classified into three classes, that is, Filter, Wrapper and Embedded, as shown in Fig. 1.

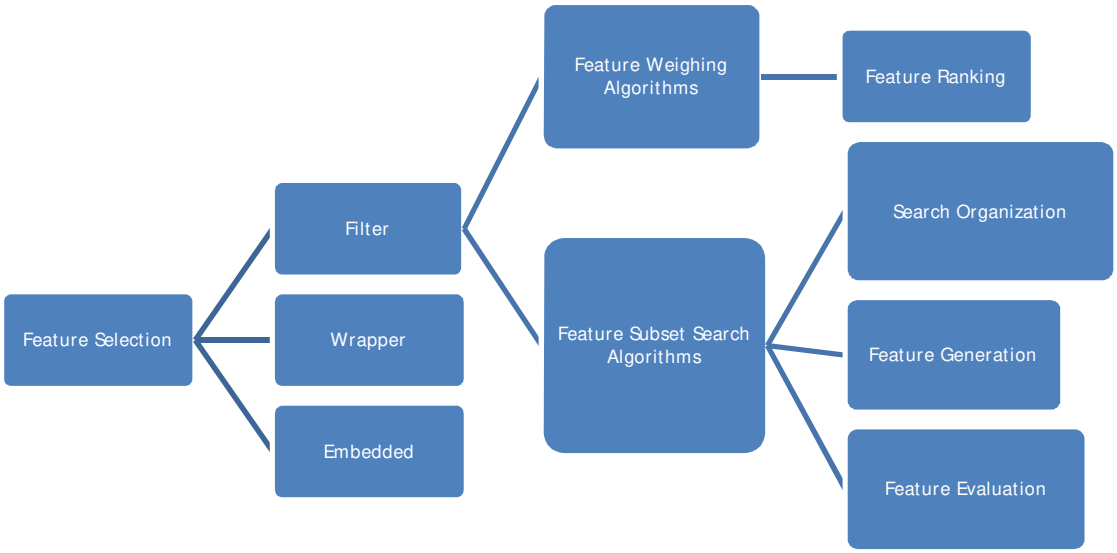


Fig. 1 Feature Selection Algorithms

A. Filter Approach

The filter function that can be univariate or multivariate depends on the characteristics of the data. The advantage of the filter approach is that they do not depend on any machine learning algorithm and are computationally simple and fast. However, the main disadvantage of this approach is that it ignores the dependencies between the characteristics and does not interact with the classifier. The filter methods used in this paper are discussed below.

1) *ReliefF:* ReliefF is a feature selection technique based on the filter that was first proposed by Kira and Rendell in 1992. It is a random selection technique that evaluates the characteristics based on the proportions of near and near hit. Relief uses Equation 1 to update the weights of the attributes:

$$W_x = W_x - \frac{D(x.r.h)}{m} + \frac{D(x,r,m)}{m},$$

When W_x symbolizes the weight of an attribute x, r is a randomly sampled instance, m is the sum of randomly sampled instances, H is the closest hit, and M is the closest error. D represents the function difference that is used to calculate the variance of an attribute for two different instances. The initial version of Relief was limited to only two class labels, but the Relief algorithm proposed by (Kononenko, 1994), which is an improvement on Relief, takes into consideration labels of various classes and also deals with noisy or incomplete data. The advantage of the Relieve algorithm over other techniques is that they use much less time, but the limitation is that redundant functions are not eliminated.

2) *Chi Squared Filter:* This is a statistical learning technique that is applied to test the dependence of two variables. In the selection of characteristics, Chi Square is used to identify the relationship between a specific characteristic and the target class and to conclude if they are dependent or independent of each other. The chi-square χ2 is commonly used for testing relationships between categorical variables [8]. The statistic χ ^ 2 for the set of variables k is:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

3) *Correlation-Based Feature Selection:* It is based on the principle that "an excellent characteristic is one that has a high positive correlation with the output class and has no correlation or is not correlated with any other characteristic of that class". To measure the degree of uncertainty of a variable, Entropy is used, which is calculated as follows:

$$H(Y) = -\sum P(y_i) \log_2(P(y_i))$$

In equation 4 above, P (yi) represents the foregoing probabilities for all values of Y, and P ($\frac{y_i}{z_j}$) are the last probabilities of Y, given the values of Z.

Information gain is calculated as:

$$IG\left(\frac{Y}{Z}\right) = H(Y) - H\left(\frac{Y}{Z}\right)$$

From this, we can conclude that, if

$$IG\left(\frac{Y}{Z}\right) > IG\left(\frac{Znew}{Z}\right), \qquad (6)$$

then the characteristic Z is extremely correlated with the characteristic Y instead of the characteristic ***Znew***.

B. Wrapper Approach

The Wrapper model depends on a specific classifier to evaluate the quality of the selected functions. Start by looking for the technique in the space of probable subsets of characteristics and creating several subsets of characteristics. The selected characteristics are used to evaluate the performance of a predefined classifier. This process is repeated until the anticipated quality is reached. The techniques based on the wrap approach used in this research are discussed below.

1) *Hill Climbing:* The hill climbing technique initially considers a random set of attributes. The neighbors of the set are evaluated and the best one is chosen. The advantage of using this technique is that it entails fewer conditions compared to others and is very useful for solving pure optimization problems. However, it afflicts with the problem of Local Maxima and Plateau. Maxima local says that "a state is better than all its neighbors, but not necessarily with those states that are far away." In mathematical terms, for a function f (x), ∀x∈R (a, f (x)) is a local maximum if there is an interval (y, z) with

$$y < x < zandf(a) \geq f(\boldsymbol{x}) \forall \boldsymbol{x} \in \boldsymbol{R}$$

2) *Random Forest:* Random forest is a very efficient algorithm introduced by Breiman in 2001. It evaluates subsets of characteristics by verifying the performance quality of the subset in a modeling algorithm. Random forest uses tree-based strategies where the nodes that have the least impurities are established as the initial nodes of the tree and the nodes with the highest impurities are established at the end of the trees. Impurity is calculated using Gini impurity or information gain. Let A be the set of all the attributes.
3) *Best First Search:* To select the subset of features, you can choose forward or backward selection techniques. Best first search proposed by P.M. Marendra and K. Funkunaga (1977) is a slight variation that is similar to the advanced search technique beyond the truth, since it selects the best of the characteristics and then evaluates it.

C. Embedded Approach

In an embedded method, feature selection is integrated or built into the classifier algorithm. During the trail, the internal are adjusted the classifier and determines the appropriate weights/importance given for each feature to produce the best classification accuracy. Therefore, the search for the optimum feature subset and model construction in an embedded method is combined in a single step (Guyon and Elisseeff, 2003). Some examples of embedded methods include decision tree-based algorithms (e.g., decision tree, random forest, gradient boosting), and feature selection using regularization models (e.g., LASSO or elastic net).

III. RESEARCH EXPERIMENT ORGANIZATION

The organization of research experimentation presented in the paper is as follows:

1) Creation of Dataset
2) Feature Selection using Filter and Wrapper Methods
3) Feature Evaluation using SVM

A. Dataset

To conduct the experiments, data was collected from graduate and postgraduate students at private universities through a structured questionnaire consisting of stuck questions related to personal, socioeconomic and non-cognitive factors. These factors are taken into consideration based on previous research and related work in this field as discussed in Section 2. The questionnaire was based on the Likert scale. The advantage of using Likert scales is that they not only have a simple yes / no answer from the respondent, but that the responder has certain degrees of opinion and even no opinion. Reliability was measured using the Cronbach's Alpha (internal consistency measure) in SPSS. The alpha value of 0.8 to 0.9 is considered good, which in our case was 0.880. Table 2 summarizes the initial set of characteristics used for the experimentation. The data set consists of 385 instances with 22 characteristics, of which 21 were independent variables and one is, "GPA" was a dependent variable that had three classes that represent the students' grade, that is, A, B or C.

B. Feature Selection using Filter and Wrapper Approach

To carry out the experiment, we used R Tool, which is an open source machine learning software for statistical computing and visualization of graphics. It is a complete programming language that has a large set of predefined function libraries. Here, we used the "FSelector" library, which has numerous predefined function selection algorithms available.

After applying several algorithms as discussed in Section 2, the ten main characteristics (academic self-efficacy, social anxiety, academic motivation, pursuit of achievements, study habits, consciousness, age, extraversion, academic stress and intrinsic motivation) and their weights they are specified in Fig. 2. In this, the characteristics are based on RF, Chi square and Relief.
Gender, Age, Extrinsic Motivation, Academic Motivation, Consciousness, Academic Self-Efficacy, Self-Discipline, Academic Stress, Assertiveness were the best features selected using Hill Climbing and Best First Search, while CFS Filter.

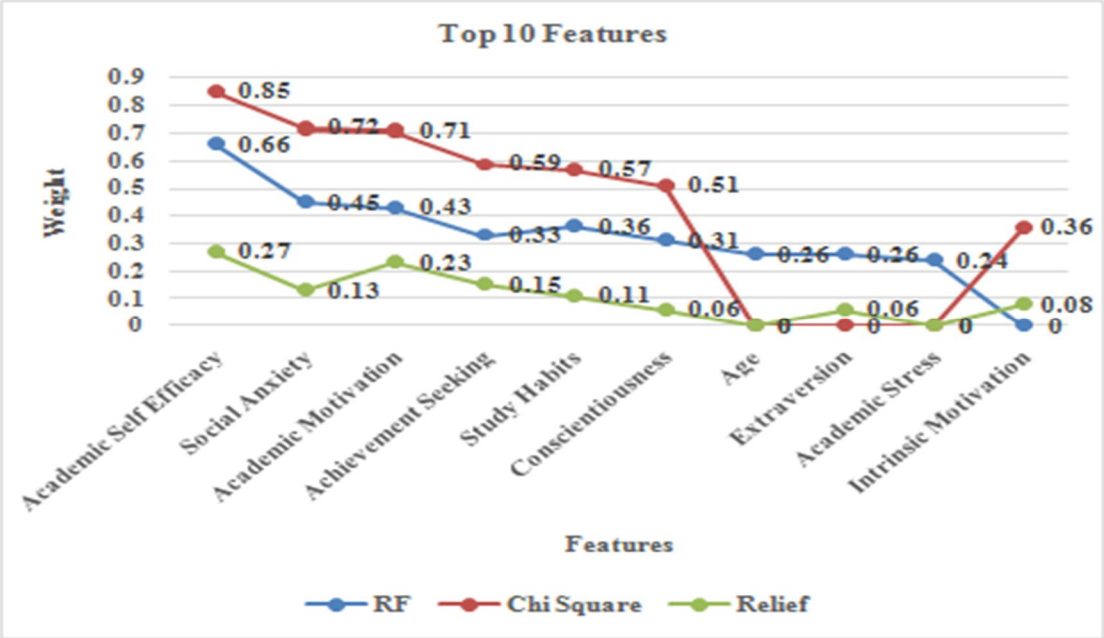


Fig. 2: Top 10 Features Ranking using RF, ReliefF and Chi-Square

Selection that uses Correlation and Entropy measures to select the best characteristics identified as Social Anxiety, Academic Motivation, Academic Self-Efficacy, Study Habits and Achievements as important attributes that affect the academic performance of students

C. Feature Evaluation

Data classification is one of the common tasks used in machine learning. The Support Vector methods was proposed by V.Vapnik in 1965, when he was trying to solve problems in pattern recognition [9] . To evaluate the performance of the model created with the selected functions, we have used Support Vector Machine (SVM), which is the most used classification technique in Data Mining. This is used to classify the students' performance according to their GPA in three classes, that is, "A", "B" or "C". The researchers in their article have confirmed that SVM Classifier is well known for its maximum accuracy. Here, to carry out the experimentation and evaluate the selected characteristics, we have divided the whole data set into two sets, that is, training data and tests. The model is trained on 70% data and the remaining 30% is used to test the model.
There are a total of 385 instances, of which approximately 270 are used to train the model and the remainder 115 is used to test the validity of the model. If the accuracy of the model is above the minimum threshold, the characteristics are selected, otherwise the process continues until the desired level is not reached.

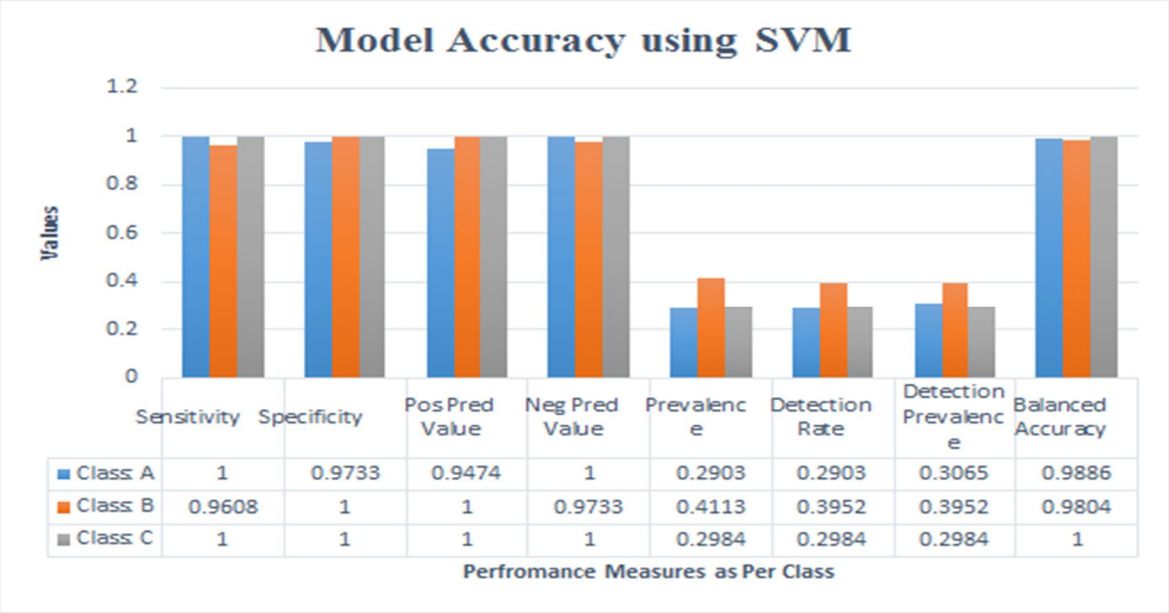
IV. INTERPRETATION OF RESULTS

To evaluate the performance of model using optimal set of features, accuracy of SVM classifier is estimated.

A. Overall Statistics

This model was constructed using the ten main characteristics as shown in Fig. 2 above. In addition, depending on the class, the accuracy of the model was also calculated, which showed that the accuracy of class C exceeded class A and B statistics.

B. Statistics by Class



	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Prevalence	Detection Rate	Detection Prevalence	Balanced Accuracy
Class: A	1	0.9733	0.9474	1	0.2903	0.2903	0.3065	0.9886
Class: B	0.9608	1	1	0.9733	0.4113	0.3952	0.3952	0.9804
Class: C	1	1	1	1	0.2984	0.2984	0.2984	1

Fig 3: Accuracy of Model as per Class using SVM

V. CONCLUSION AND FUTURE SCOPE

The present research emphasizes the identification, selection and evaluation of several characteristics that affect the academic performance of students at the undergraduate and graduate levels. There are many defensible reasons that support the evaluation of these characteristics, for example, an incorrect evaluation can result in a low quality education. In the present study, a total of 22 psychological factors related to the academic performance of the students in the initial stage were identified, then techniques of feature selection (filter and wrapping technique) were implemented to select the most influential characteristics. Finally, 10 characteristics are selected to carry out the additional evaluation using the SVM classifier approach. SVM classifier is currently more popular classifier [10]. The current work can be improved considering other existing classifiers such as Naive Bayes, Decision Tree, k-Nearest Neighbor, etc. that will help to identify the optimal generic model to achieve a high quality education.

REFERENCES

[1] D. K. Arun, V. Namratha, B. V. Ramyashree, Y. P. Jain and A. Roy Choudhury, "Student Academic Performance Prediction using Educational Data Mining," 2021 International Conference on Computer Communication and Informatics (ICCCI), 2021, pp. 1-9, doi: 10.1109/ICCCI50826.2021.9457021.

[2] A. O. Enaro and S. Chakraborty, "A Review on the Academic performance analysis of higher Education Students using Data Mining Techniques," *2019 International Conference on contemporary Computing and Informatics (IC3I)*, 2019, pp. 256-260, doi: 10.1109/IC3I46837.2019.9055632.

[3] M. S. Ram, V. Srija, V. Bhargav, A. Madhavi and G. S. Kumar, "Machine Learning Based Student Academic Performance Prediction," *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, 2021, pp. 683-688, doi: 10.1109/ICIRCA51532.2021.9544538.

[4] Hussain, S., Khan, M.Q. Student-Performulator: Predicting Students' Academic Performance at Secondary and Intermediate Level Using Machine Learning. Ann. Data. Sci. (2021). https://doi.org/10.1007/s40745-021-00341-0

[5] Yağcı, M. Educational data mining: prediction of students' academic performance using machine learning algorithms. Smart Learn. Environ. **9**, 11 (2022). https://doi.org/10.1186/s40561-022-00192-z

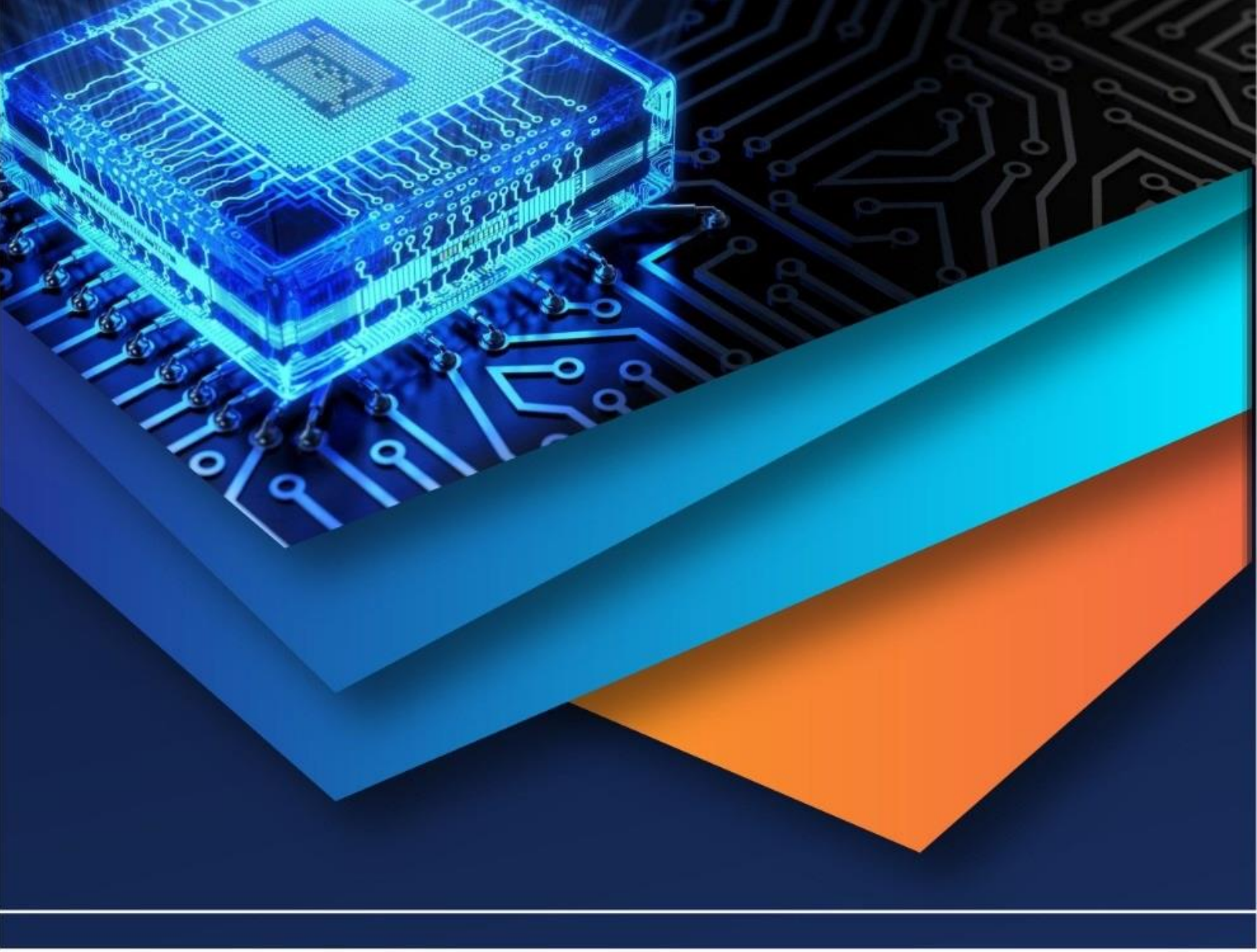
[6] Saba Bashir, Irfan Ullah Khattak, Aihab Khan, Farhan Hassan Khan, Abdullah Gani, Muhammad Shiraz, "A Novel Feature Selection Method for Classification of Medical Data Using Filters, Wrappers, and Embedded Approaches", Complexity, vol. 2022, Article ID 8190814, 12 pages, 2022. https://doi.org/10.1155/2022/8190814

[7] Abiodun, E.O., Alabdulatif, A., Abiodun, O.I. et al. A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Comput & Applic **33**, 15091–15118 (2021). https://doi.org/10.1007/s00521-021-06406-8

[8] Sarker, I.H. Machine Learning: Algorithms, Real-World Applications and Research Directions. SN COMPUT. SCI. 2, 160 (2021). https://doi.org/10.1007/s42979-021-00592-x

[9] Q. Wang, "Support Vector Machine Algorithm in Machine Learning," 2022 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), 2022, pp. 750-756, doi: 10.1109/ICAICA54878.2022.9844516.

[10] Yujun Yang, Jianping Li and Yimei Yang, "The research of the fast SVM classifier method," 2015 12th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2015, pp. 121-124, doi: 10.1109/ICCWAMTIP.2015.7493959.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY