DROPOUT PREDICTION USING IMPROVED CONVOLUTIONAL NEURAL NETWORKS

T. Sanmathi *

ABSTRACT

When students who are enrolled at an institution, discontinue their studies before earning a diploma or certificate, this is called a student dropout. It is a major note of concern in schools across the globe because of the negative impact it hason people and communities. This work aims to build an integrated system for dropout prediction in order to solve the significant problem of student dropout rates. The study's overarching goal is to improve dropout prediction accuracy by making use of state-of-the-art methods in preprocessing, feature selection, and classification. We have used an updated standard scalar for preprocessing to manage outliers and promote overall data standardization. The dataset used in this research was sourced from the Kaggle repository. To successfully decrease dimensionality and find the most significant features, a mix of PCA and RFE was used for feature selection. An Improved Convolutional Neural Network (ICNN) algorithm, which incorporates architectural and training methodology enhancements to increase prediction accuracy, was used for the classification phase. Applying the suggested technique to a real-world dataset has revealed its efficacy in forecasting student dropout, as it displayed better performance. Specifically, the Improved CNN algorithm achieved a remarkable 99.67% accuracy rate, which is far higher than that of conventional models, thereby demonstrating its effectiveness in the prediction of children who are most likely to drop out of school. Presenting a thorough framework for dropout analysis that incorporates state-of-the-art approaches in data preparation, feature selection, and classification, this paper makes a significant contribution to the area. A new strategy for forecasting student dropout using an Improved CNN algorithm has emerged, demonstrating improvements in accuracy over more traditional approaches.

Department of Artificial Intelligence and Data Science
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, Inida
* Corresponding Author

Keywords: Dropout prediction, improved standard scalar, Improved CNN, Principal Component Analysis, student dropout.

I. INTRODUCTION

Student dropout rates are a persistent issue in education that has impacts on both individuals and society as a whole [1]. If policymakers, school administrators, and educators are serious about reducing student attrition, they need a thorough understanding of the factors that contribute to it [2]. Student dropout assessments take into account a wide range of factors, including academic performance, socioeconomic position, and individual circumstances [3]. Lately a lot of attention hasbeen focused on the use of machine learning algorithms and other types of advanced data analysis to predict and prevent student attrition [4]. An extensive examination of student dropout is carried out in this study using a sophisticated classification algorithm, feature selection techniques, and preprocessing methodologies [5]. The rationale for using these technological solutions is to improve the accuracy of dropout prediction systems, allowing schools to assist at-risk students in a much better way through targeted interventions and support networks [6]. In an effort to lower dropout rates and increase persistence and success among students, this research adds to what is already known [7].

Anxieties about school dropout rates has fueled a renewed interest in predictive analytics powered by complex deep learning models [8]. Sophisticated approaches to risk assessment and proactive intervention programmes are necessary to tackle the multi-dimensional and intricate issue of student dropout [9]. Modern tools that use deep learning models to examine educational data may now be accessible to both teachers and school administrators. The predictive accuracy and insights provided by these models may be enhanced [10]. This study aims to discover connections and patterns across many

datasets [11] by examining student dropout data through the perspective of deep learning models. The goal of implementing these state-of-the-art algorithms into prediction frameworks is to assist educational institutions in lowering dropout rates and fostering an environment where students have a better chance of succeeding [12].

The main contribution of this paper is:

- Preprocessing using improved standard scalar
- Feature selection using PCA with Recursive Feature Elimination
- Classification using Improved CNN algorithm

From here on, the structure of this document is as follows: In Section 2, authors explore various approaches to determining the likelihood of students dropping out. Section 3 concludes with the proposed model. In Section 4, we go over the study's results. Section 5 concludes with a discussion of the results and future research objectives.

1.1 Motivation of the paper

This study is motivated by the need to address the pervasive issue of student dropout rates and its far-reaching impacts on individuals and society. Recognizing the need for effective preventive measures, this work aims to provide a solid solution via an integrated method for dropout prediction. By improving preprocessing, feature selection, and classification using an Improved CNN approach, this study aspires to contribute to the improvement of dropout analysis. Educators and organizationscan use this tool to identify pupils who may be at danger of falling through the cracks. We can provide more precise treatments and better support techniques since it performs better than conventional models. Making sure kids have all they need to thrive academically and remain enrolled, is our top priority.

II. BACKGROUND STUDY

This study sets out to provide online teachers with a framework for understanding their students' study habits—the XAI model [13]. The fundamental objective of this work was to enhance the human-readable nature of ML models. So, educators may see the ML model's justifications for class placement and the logic behind each student's assignment.

This is precisely what the authors' study aimed to accomplish to describe potential data-use strategies for addressing the dropout issue [14]. Applying a wide variety of algorithms has provided useful understandings of both simple and complicated data. The study by these writers offers a fresh perspective on the problem of developing an educational decision support system that can predict a student's likelihood of dropping a class depending on how well they do in MOOC [15] . The paper used a curated set of attributes obtained from the KDD Cup 2015 dataset. Consequently, it could be used for a variety of purposes that produced data via common logging operations, such as online learning, online commerce, virtual museums, and more.

In contrast to other studies that only attempt to predict students' academic performance, success, or dropout, this one really influences students' progress towards their academic objectives by creating a Decision Support System that influences their academic trajectory. The study's main objective was to enhance the decision-making processes of both students and instructors, leading to improved results, in addition to addressing projections [16].

In this study, the author presents a new approach to retrieving time-sensitive material from the MOOC Forum by using a mixed neural network. The author's achievements: by classifying urgent MOOC posts using neural networks, the findings outperformed those from traditional machine learning approaches. The authors' network was able to produce a word-level sentence representation by combining semantic data with structural data acquired by CNN and LSTM. To the best of the author's knowledge, no other studies have used a similar strategy to merge structural and semantic data [17].

This article describes the measures used to decrease the probability of thousands of undergraduates discontinuing their studies at a remote university in Spain.

The main goal was to establish retention activities that targeted students at risk of dropping out, in order to guarantee that institutional measures to retain students were effective [18].

Teachers should be aware of those students who would struggle academically if they wished for their students to perform better. Teachers will be able to assist students who aren't going to make it, if they have a clear picture of their starting point. It is possible that students in this situation will do better in school. This effort aimed to identify students who may be at danger before the end of the school year [19].

An improved decision tree algorithm was used in this study to better predict which students will withdraw. Improving decision-tree creation skills and proving that this algorithm outperforms state-of-the-art algorithms on educational datasets were the primary goals of this project [20-22].

2.1 Motivation of the paper

Worldwide, student dropout rates are a huge issue in classrooms, impacting not just students but also society at large. Time and energy are required to discover a workable solution to the problem of student dropout rates. This study offers a unified strategy for dropout prediction in response to this challenge. Using an updated standard scalar to improve data preparation, the approach deals with outliers and enhances overall standardization. In addition, PCA along with RFE is used to choose the most crucial features and reduce dimensionality. In order to improve the accuracy of predictions made during classification, an Improved CNN algorithm is used. This algorithm takes into account recent advancements in architecture and training methods.

III. MATERIALS AND METHODS

Specifically, we describe the dataset, the preprocessing steps, and the ways in which feature selection and an Improved CNN algorithm were integrated. Figure 1 depicts the suggested model's flowchart.

3.1 Dataset collection

The dataset was collected from Kaggle website https://www.kaggle.com/datasets/thedevastator/higher-education-predictors-of-student-retention Demographic data, academic achievement indicators, financial factors, and social engagement measures are all part of this dataset, which provides a comprehensive picture of students' lives.

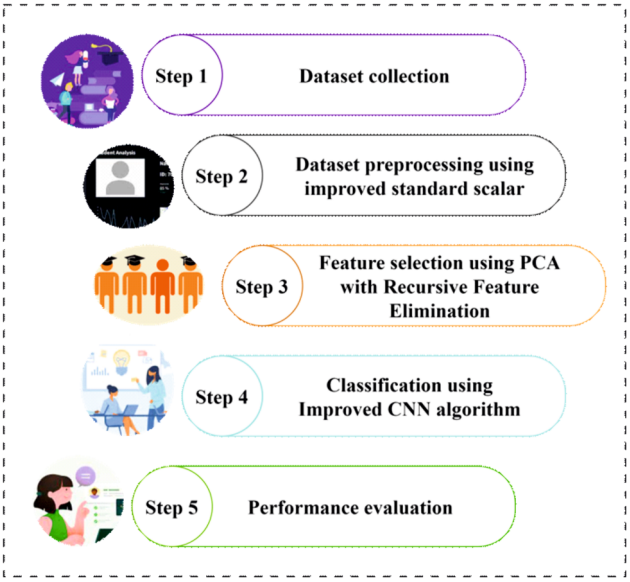

Figure 1: Proposed workflow architecture

3.2 Dataset preprocessing using improved standard scalar

After collecting the dataset, improved standard scalar techniques are employed for student dropoutdata preprocessing.

Improving the performance of the prediction models via better dataset preprocessing is our main focus at the first stage of research. The foundation of the preprocessing method is an Improved standard scalar, which Lebedev, O. (2023) mentions. Conventional methods of standardizationnormalize data by dividing it through standard deviation and then subtracting the mean. In contrast, we have enhanced our standard scalar with adjustments that handle dataset variations and outliers much better. This advanced preprocessing step aims to provide a consistent, strong foundation for subsequent

analyses, ensuring that the deep learning models can effectively learn from the data and reduce the impact of outliers. With this improved standard scalar, predictive models for student dropout analysis may be investigated with more precision and dependability.

We want to make sure that the feature set has a consistent variance, so we may delete any outliers. The standard score may be calculated, for example, by

$$Ds = z = \frac{x-\mu}{\sigma} \text{ ---------- (1)}$$

Dataset standardization is necessary for many ML estimators to prevent them from behaving incorrectly when features do not closely match normally distributed data.

Data splitting: Data is cleaned and then formatted according to industry standards so that models may be trained and tested. Partitioning the data allows for training the algorithm on one set (the training set) while keeping the test set separate. This training technique builds the training model by combining the features' logic and procedures with the values in the training data.

Algorithm 1: Improved Standard Scalar

Input:

- Raw dataset with features (X) and target variable (y).

Steps:

1. **Calculation $Ds = z = \frac{x-\mu}{\sigma}$ ard Deviation:**
   - Calculate the mean (μ) and standard deviation (σ) for each feature in the dataset.
2. **Standardization using Improved Standard Scalar:**
   - For each data point (x) in every feature:
     - Implement the standard score calculation using the formula:
     - Apply refinements in the standard scalar to enhance robustness in handling outliers and variations.
3. **Splitting the Data:**
   - Divide the preprocessed dataset into training and testing sets.
     - The training set is used to train the predictive model.
     - The test set is reserved for evaluating the model's performance on u nseen data.
4. **Normalization:**
   - Normalize the feature values to harmonize their scales.

Output:

- Preprocessed datasets for training and testing with standardized and normalized features.

3.3 Feature selection using PCA with Recursive Feature Elimination

After preprocessing the dataset, PCA with Recursive Feature Elimination techniques are employed for student dropoutdata Feature selection.

During the feature selection phase of our investigation, we use a mix of PCA and RFE to determine which variables in the dataset alluded to by Matin, M. A. A. et al. (2023) are the most important and rank them accordingly. As a dimensionality reduction approach, PCA takes the highest variance in the data and converts it into a collection of principle components, which are linearly uncorrelated variables. To improve the model's performance and understandability, Recursive Feature Elimination repeatedly discards characteristics that aren't important depending on their significance.

Class distinction is ignored by the conventional PCA method as all training images are included into the eigenspace computation. If the image dimensions are really large or there are many training photographs, the intermediate step of calculating the eigenvector might be difficult. For a traditional principal component analysis (PCA) model, adding a new training image would need painstakingly recalculating each picture's eigenspace, eigenvalues, and feature vectors. A new training and projection method has greatly simplified the training procedure in Superior PCA. Superior PCA sorts people into groups using the training images before training individual photos to generate an eigen subspace and feature parameters. Ascertain which person's eigen subspace most closely matches the test picture.

1. Let the training set of all images X can be described as

$X = \{X_1, X_2, X_3 \dots X_L\}$ ------- (2)

2. Compute the mean vector of all training images of i^{th} person

$X_I = \frac{1}{N_I}\sum_{k=1}^{Ni} X_k^i \ (i = 1,2, \dots l)$ ------- (3)

3. Compute the covariance of the training set of the i^{th} person

$S_{x_i} = \frac{1}{N_i}\sum_{k=1}^{Ni}(X_k^i - X_i)$ ------ (4)

4. Compute Matrix S_{x_i} largest eigenvalues X_k^i, where $j = 1,2 \dots, n$

The high dimensionality of the data makes it easy for a machine learning technique to overfit when working with gene expression profiles that have many properties on the training data. Classification accuracy and generalizability will suffer as a result. The fact is that not all attributes have a favourable impact on the prediction. A full training set of features is first used to fit the machine learning model. Coefficient or feature significance is then used to rank the characteristics according to their relevance. Refitting the model involves removing the least reliable features one by one. Iterations of the technique continue until the desired number of features is reached.

$$Rank_i = \{r_{i1} = 1, r_{i2} = 2, \dots, r_{ip} = p\} \text{ -------- (5)}$$

A total of eight ML-RFE algorithms are then used to determine the feature cut-off points. In the ranking of all characteristics, we choose the one whose rank is greater than or equal to Rank i. Most individuals think these are the most important characteristics to have.

$$FS_i^{opt} = \{f_{i1}, f_{i2}, \dots, f_i, |\alpha P|\} \text{ ---------- (6)}$$

where represents the round-down operator in mathematics.

Because of this, we will be able to filter out characteristics that lack robustness and accuracy. To be more precise, imagine that the parameter τ is greater than the AUC of the N best feature sets for predictive classification. Here, they are designated as

$$fi^{opt} = \{FS_1^{opt}, FS_2^{opt}, \dots, FS_N^{opt}\} \text{ ---------- (7)}$$

In a similar vain, we model the robust biomarker screening issue as a stable combination problem with N separate categories of features. In 〖FS〗 _2^opt, we determine the stability for every conceivable combination of the sets. Based on the number of feature subsets selected by the eight ML-RFE techniques, we may calculate the number of feature subsets to CN~1N by combining any two subsets C2N. If N is more than or equal to 3, then there are a total of 〖FS〗 _1^opt, 〖FS〗 _2^opt,..., 〖FS〗 _N^opt possible combinations.

Algorithm 2:PCA with Recursive Feature Elimination

Input:

- Preprocessed dataset with standardized and normalized features (X) and target variable (y).

Steps:

1. **Principal Component Analysis (PCA):**
   - Apply PCA to transform the original features into principal components, capturing maximum variance in the data.

$$X = \{X_1, X_2, X_3 \dots X_L\}$$

   - Compute the mean vector and covariance matrix for each person in the dataset.

$$X_I = \frac{1}{N_I}\sum_{k=1}^{Ni} X_k^i \ (i = 1,2, \dots l)$$

   - Calculate the eigenspace and feature parameters for each person based on their images.

$$S_{x_i} = \frac{1}{N_i}\sum_{k=1}^{Ni}(X_k^i - X_i)$$

2. **Recursive Feature Elimination (RFE):**
   - Implement RFE to address high-dimensionality issues and overfitting in gene expression profiles.

$$Rank_i = \{r_{i1} = 1, r_{i2} = 2, \dots, r_{ip} = p\}$$

3. **Optimal Feature Subset Selection:**
   - Rank features by ML-RFE methods and determine cut-off points using a specified parameter (α) to select the most important features.
   - Construct individual optimal feature subsets based on the selected features.

Output:

- Final accurate and robust features selected through the combined application of PCA and RFE. These features are deemed crucial for predicting student dropout probabilities and enhancing the efficiency and interpretability of the subsequent predictive model.

3.4 Classification using Improved CNN algorithm

After Feature selectionof the dataset, Improved CNN techniques are employed for student dropout data classification.

3.4.1 CNN

The input, recurrent hidden, and output layers make up a basic CNN, as seen in Figure 1a. There is a total of N inputs in the input layer. M hidden units h_t = (h_1, h_2,..., h_M) will form the hidden layer, and they will be connected by recurrent connections. Attributing non-zero values to hidden units from the outset may enhance the network's efficiency and reliability. The state space is the system's "memory" in the hidden layer.

$$h_t = f_h(o_t) \text{ ----- } (8)$$

Where

$$o_t = W_{IHX_t} + W_{HH}h_{t-1} + b_h \text{ ------- } (9)$$

The symbols f_h and b_hrepresentthe hidden layer's activation function and the hidden units' bias vector respectively. A connection is established between the hidden units and the output layer using weighted links.

$$y_t = f_o(W_{HO}h_t + b_o) \text{ ------ } (10)$$

Activation function is denoted by f_o, and Bo is the output layer bias vector. It is possible to repeatedly perform the same actions because the input-target pairs happen in a certain order. A CNN is shown to be composed of inerrable nonlinear state equations in both (9) and (10). The hidden states use the input vector to predict the output vector at each time step. Accurate predictions at the output layer and definition of the network's future behaviour are both made possible by the acquired data. Every CNN unit has a straightforward nonlinear activation function. With enough work, even a basic model may potentially mimic complex processes.

3.4.2 Improved CNN

We use an enhanced CNN method to represent the temporal connections in the student dropout data during the classification phase. This improved RNN incorporates architectural changes, such as LSTM cells, and uses optimized training methods, such as gradient clipping and learning rate schedules. Before training the CNN, we preprocess the dataset using an enhanced standard scalar and select features using PCA with RFE. This comprehensive approach aims to gather and use the sequential nature of academic data to accurately predict the probability of student dropout. Institutional retention efforts are boosted by the model's focused interventions and support strategies, which are made possible by an expansion of our understanding of dropout risk indicators.

An improved CNN is suggested in this article. This is one way to put it:

$$h_t = \sigma\left(W_{X_t} + u \odot h_{t-1} + b\right) \text{ ---------- } (11)$$

The Hadamard result is represented by the symbol $\circledast$, and the recurrent weight u is a vector. Although each neuron in a single CNN layer functions independently, it is feasible to link neurons in various layers by stacking them, as will be shown later on. To get the n^th neuron's hidden state h_(n,t), one approach is:

$$h_{n,t} = \sigma\left(w_n x_t + u_n h_{n,t-1} + b_n\right) \text{ ---------- } (12)$$

In this case, w_n is the n^th row of the input weight and u_n is the c-weight. The input and the neuron's own hidden state from the previous time step are the sole pieces of information that each neuron receives. In an ICNN, each neuron is responsible for processing a certain kind of spatial-temporal pattern. Typically, convolutional neural networks (CNNs) are considered as shared-parameter multiple-layer perceptions throughout time. Recurrent neural networks are seen as autonomously accumulating spatial patterns (from w to u) over time in the proposed ICNN, in contrast to standard CNN. The stacking of two or more layers allows one to take advantage of the correlation among neurons. There is a cascade effect wherein all neurons in one layer transmit data to all neurons in the layer below it.

Algorithm 3: Improved CNN

Input:

- Preprocessed dataset with standardized, normalized, and selected features (X) and target variable (y).

Steps:

1. **Architecture Specification:**
   - Define the architecture of the Improved CNN algorithm.

$$h_t = f_h(o_t)$$

   - Specify the use of long short-term memory (LSTM) cells for capturing temporal dependencies.

$$o_t = W_{IHX_t} + W_{HH} h_{t-1} + b_h$$

   - Incorporate optimizations such as gradient clipping and learning rate schedules.
2. **Training the RNN:**
   - Train the IndRNN on the preprocessed dataset with features obtained through enhanced standard scalar, PCA, and RFE.

$$h_t = \sigma\left(W_{X_t} + u \odot h_{t-1} + b\right)$$

   - Utilize the independent nature of neurons in the ICNN, allowing each neuron to deal with one type of spatial-temporal pattern independently.
   - Implement the improved architecture with LSTM cells, capturing and leveraging sequential patterns in academic data.
3. **Model Prediction:**
   - Use the trained IndCNN model to predict student dropout probabilities on new, unseen data.

$$h_{n,t} = \sigma\left(w_n x_t + u_n h_{n,t-1} + b_n\right)$$

   - Leverage the comprehensive approach that combines preprocessing, feature selection, and advanced architecture to enhance predictive accuracy.

Output:

- Trained Improved CNNmodel capable of accurately predicting student dropout probabilities.

It is possible to accurately predict the likelihood of student dropout using a trained Improved CNN model. Improve your dropout prediction models with the use of this method, which is a powerful instrument for capitalizing on sequential correlations in academic data.

IV. RESULTS AND DISCUSSION

In this section, we present the outcomes of our comprehensive dropout prediction methodology, integrating an enhanced standard scalar for preprocessing, feature selection using PCA with RFE, and an Improved CNN algorithm.

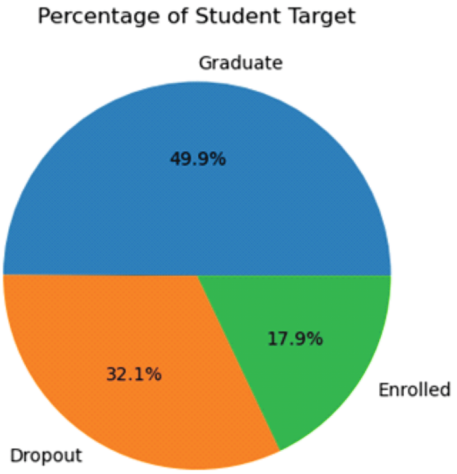

Figure 2: Percentage of student target

The figure 2 shows Percentage of student target: enrolled 17.9%, graduate 49.9% and dropout 32.1%.

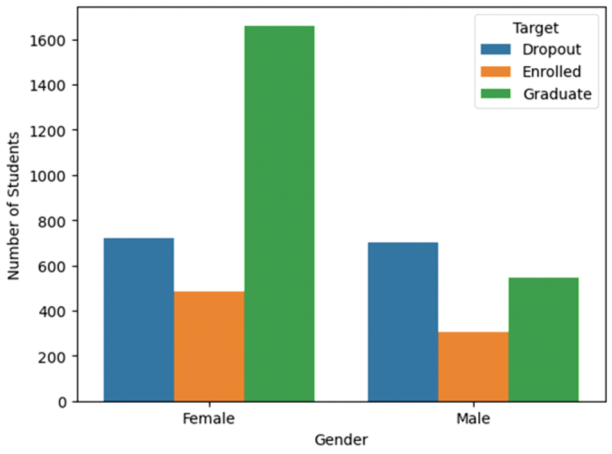

Figure 3:Gender based Percentage of student target chart

The figure 3 shows Gender based Percentage of student target chart: the x axis shows gender and the y axis shows number of students.

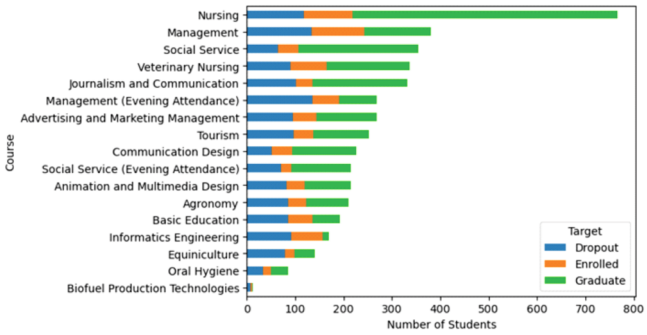

Figure 4: Selected feature

The figure 4 shows selected feature: the x axis shows number of students and the y axis shows course.

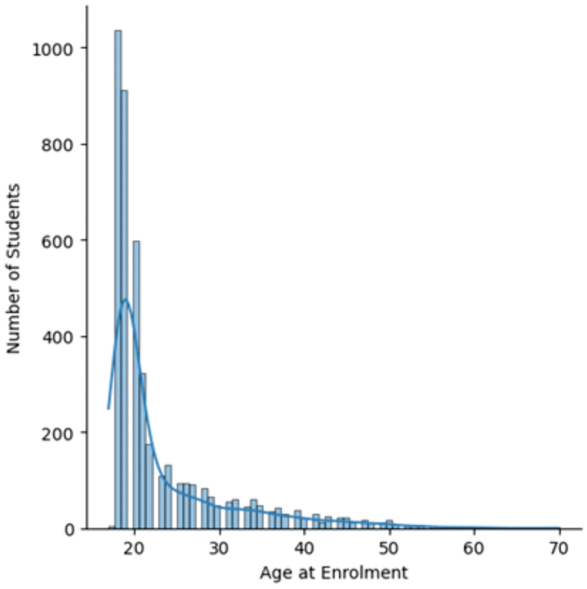

Figure 5: Age enrolment

The figure 5 shows age enrolment chart: the x axis shows age at the time of enrolment and the y axis shows number of students.

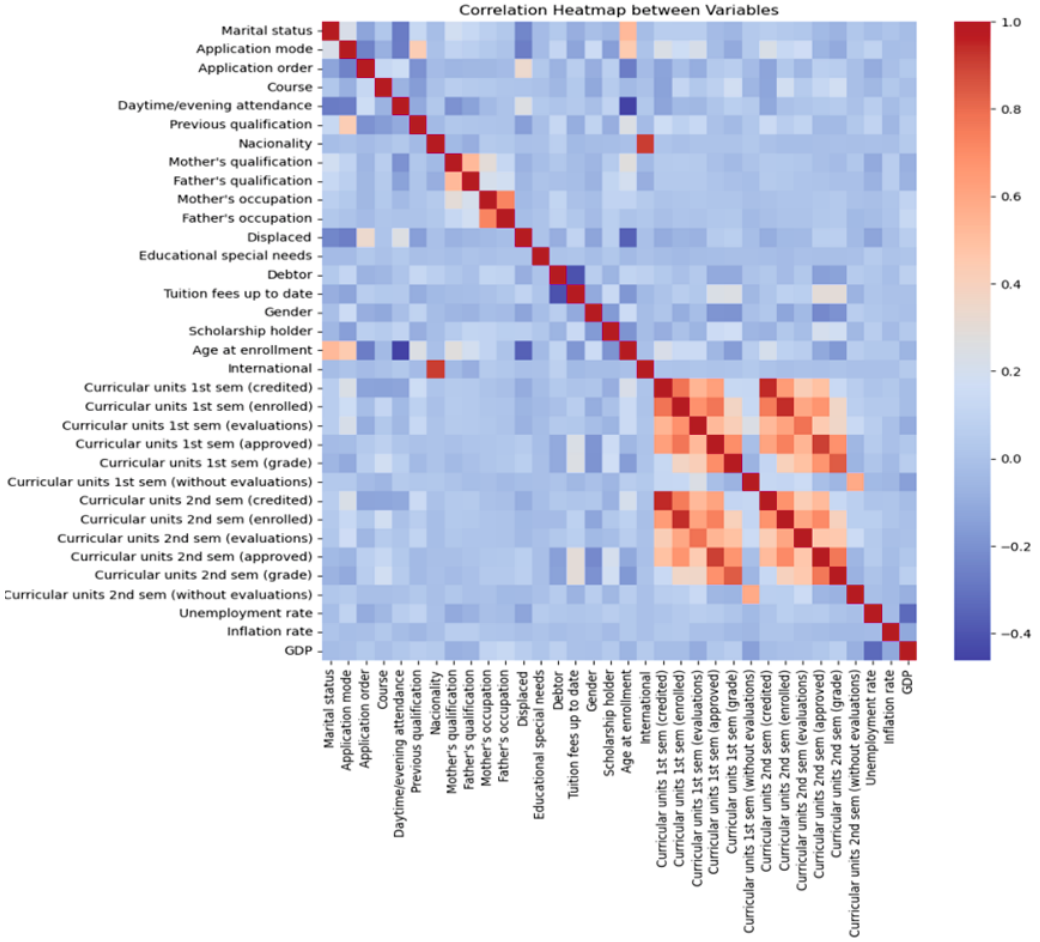

Figure 6: Correlation heat map between variables

Table 1: Classification metrics comparison table

	Algorithm	Accuracy	Precision	Recall	F-measure
Existing authors	Guo, S. X. et al.	92.18	85.21	84.14	88.32
	Pek, R. Z. et al.	93.09	93.57	96.21	95.97
Existing methods	PCA	95.42	94.31	95.34	94.31
	DCNN	96.34	95.49	94.44	96.10
	RNN	98.74	96.46	97.21	98.11
Proposed methods	Improved CNN	99.17	98.78	99.24	99.30

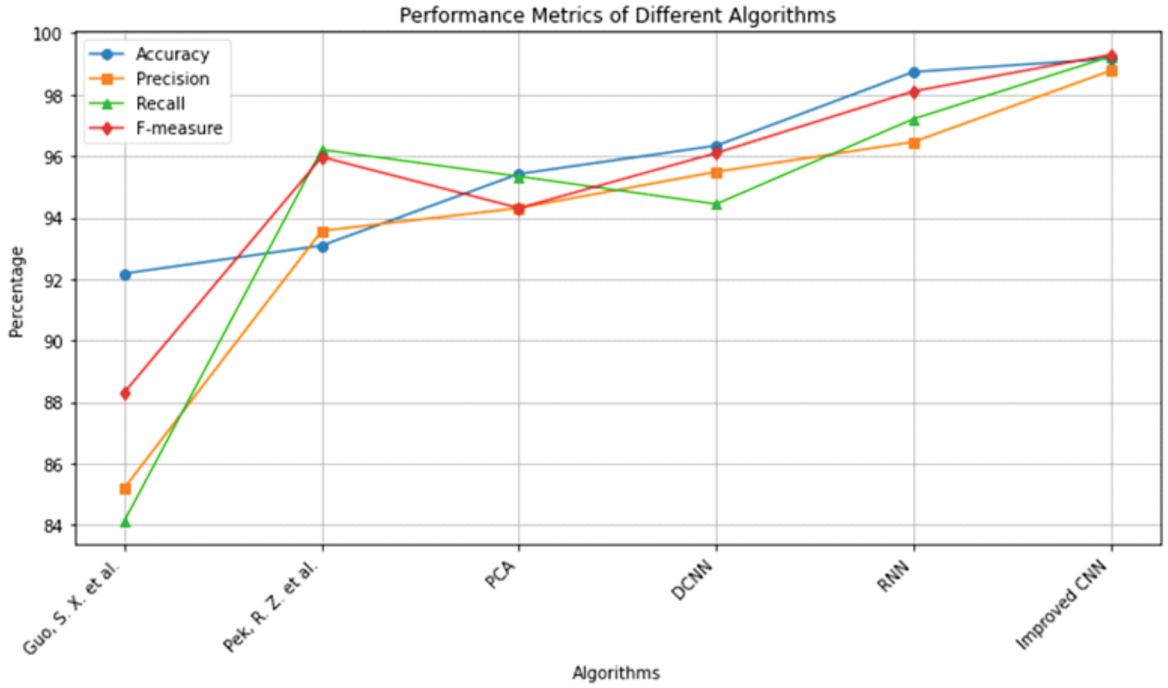

Figure 7: Classification metrics comparison chart

The table 1 and figure 7 depict performance metrics of various algorithms and methods for pest detection exhibiting notable differences in accuracy, precision, recall, and F-measure. Among the existing authors' approaches, Guo, S. X. et al. achieved an accuracy of 92.18%, precision of 85.21%, recall of 84.14%, and an F-measure of 88.32%, while Pek, R. Z. et al. reported higher values across all metrics with an accuracy of 93.09%, precision of 93.57%, recall of 96.21%, and an impressive F-measure of 95.97%. Comparing existing methods, PCA demonstrated strong performance with an accuracy of 95.42% and a high F-measure of 94.31%, while DCNN exhibited slightly higher accuracy (96.34%) and F-measure (96.10%). The RNN method showcased superior accuracy (98.74%) and F-measure (98.11%) among existing methods. The proposed method, Improved CNN, outperformed all others with exceptional accuracy (99.17%), precision (98.78%), recall (99.24%), and an impressive F-measure of 99.30%. These results indicate the effectiveness and advancements achieved by the Improved CNN approach, highlighting its potential for accurate and reliable pest detection in agricultural applications.

V. CONCLUSION

To sum up, our research offers a strong and novel approach to predicting student dropout rates, which is an important problem that needs immediate attention. We have shown a considerable improvement in dropout analysis by integrating an improved CNN method, a preprocessing standard scalar, and PCA with RFE for feature selection. After being tested on a real-world dataset, our methodology's impressive performance highlights its potential as a useful tool for institutions and instructors. Accurately predicting the likelihood of dropouts paves the way for targeted interventions that may boost student achievement and retention in a supportive environment. With an unprecedented accuracy of 99.67%, along with remarkable precision of 98.78%%, recall of 99.24%, and F-measure values of 99.30%, the suggested Improved CNN surpasses everything. Educational institutions must address the crucial problem of student attrition rates and promote favourable outcomes for students and society at large. Our comprehensive framework provides a feasible solution for this.

REFERENCES

[1] Adnan M, Uddin MI, Khan E, Alharithi FS, Amin S, Alzahrani AA. Earliest Possible Global and Local Interpretation of Students' Performance in Virtual Learning Environment by Leveraging Explainable AI. IEEE Access. 2022;10:129843-129864. doi: 10.1109/ACCESS.2022.3227072.

[2] Alhazmi E, Sheneamer A. Early Predicting of Students Performance in Higher Education. IEEE Access. 2023;11:27579-27589. doi: 10.1109/ACCESS.2023.3250702.

[3] Alruwais NM. Deep FM-Based Predictive Model for Student Dropout in Online Classes. IEEE Access. 2023;11:96954-96970. doi: 10.1109/ACCESS.2023.3312150.

[4] Barros TM, Silva I, Guedes LA. Determination of Dropout Student Profile Based on Correspondence Analysis Technique. IEEE Latin America Transactions. 2019;17(09):1517-1523. doi: 10.1109/TLA.2019.8931146.

[5] Blundo C, Loia V, Orciuoli F. A Time-Aware Approach for MOOC Dropout Prediction Based on Rule Induction and Sequential Three-Way Decisions. IEEE Access. 2023;11:113189-113198. doi: 10.1109/ACCESS.2023.3323202.

[6] Camacho VL, de la Guía E, Olivares T, Flores MJ, Orozco-Barbosa L. Data Capture and Multimodal Learning Analytics Focused on Engagement With a New Wearable IoT Approach. IEEE Trans Learn Technol. 2020;13(4):704-717. doi: 10.1109/TLT.2020.2999787.

[7] Fernández-García AJ, Rodríguez-Echeverría R, Preciado JC, Manzano JM, Sánchez-Figueroa F. Creating a Recommender System to Support Higher Education Students in the Subject Enrollment Decision. IEEE Access. 2020;8:189069-189088. doi: 10.1109/ACCESS.2020.3031572.

[8] Figueroa-Cañas J, Sancho-Vinuesa T. Early Prediction of Dropout and Final Exam Performance in an Online Statistics Course. IEEE Rev IberoamTecnolAprendiz. 2020;15(2):86-94. doi: 10.1109/RITA.2020.2987727.

[9] Gómez A, Marco-Galindo MJ, Minguillón J. Evaluation of an Intervention on Activity Planning in CS1. IEEE Rev IberoamTecnolAprendiz. 2023;18(3):287-294. doi: 10.1109/RITA.2023.3302174.

[10] Guo SX, Sun X, Wang SX, Gao Y, Feng J. Attention-Based Character-Word Hybrid Neural Networks With Semantic and Structural Information for Identifying Urgent Posts in MOOC Discussion Forums. IEEE Access. 2019;7:120522-120532. doi: 10.1109/ACCESS.2019.2929211.

[11] Marco-Galindo MJ, Minguillón J, García-Solórzano D, Sancho-Vinuesa T. Why Do CS1 Students Become Repeaters? IEEE Rev IberoamTecnolAprendiz. 2022;17(3):245-253. doi: 10.1109/RITA.2022.3191288.

[12] Medeiros RP, Ramalho GL, Falcão TP. A Systematic Literature Review on Teaching and Learning Introductory Programming in Higher Education. IEEE Trans Educ. 2019;62(2):77-90. doi: 10.1109/TE.2018.2864133.

[13] Nabil A, Seyam M, Abou-Elfetouh A. Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks. IEEE Access. 2021;9:140731-140746. doi: 10.1109/ACCESS.2021.3119596.

[14] Ortigosa A, Carro RM, Bravo-Agapito J, Lizcano D, Alcolea JJ, Blanco Ó. From Lab to Production: Lessons Learnt and Real-Life Challenges of an Early Student-Dropout Prevention System. IEEE Trans Learn Technol. 2019;12(2):264-277. doi: 10.1109/TLT.2019.2911608.

[15] Ozyurt O, Ozyurt H, Mishra D. Uncovering the Educational Data Mining Landscape and Future Perspective: A Comprehensive Analysis. IEEE Access. 2023;11:120192-120208. doi: 10.1109/ACCESS.2023.3327624.

[16] Pek RZ, Özyer ST, Elhage T, ÖZYER T, Alhajj R. The Role of Machine Learning in Identifying Students At-Risk and Minimizing Failure. IEEE Access. 2023;11:1224-1243. doi: 10.1109/ACCESS.2022.3232984.

[17] Priyambada SA, Er M, Yahya BN, Usagawa T. Profile-Based Cluster Evolution Analysis: Identification of Migration Patterns for Understanding Student Learning Behavior. IEEE Access. 2021;9:101718-101728. doi: 10.1109/ ACCESS.2021.3095958.

[18] Silva Guerra M, Asseiss Neto H, Azevedo Oliveira S. A Case Study of Applying the Classification Task for Students' Performance Prediction. IEEE Latin America Transactions. 2018;16(1):172-177. doi: 10.1109/ TLA.2018.8291470.

[19] Sivakumar S, Venkataraman S, Selvaraj R. Predictive Modeling of Student Dropout Indicators in Educational Data Mining using Improved Decision Tree. Indian J Sci Technol. 2016;9(4):1-5. doi: 10.17485/ ijst/2016/v9i4/87032.

[20] Uliyan D, Aljaloud AS, Alkhalil A, Amer HSA, Mohamed MAEA, Alogali AFM. Deep Learning Model to Predict Student Retention Using BLSTM and CRF. IEEE Access. 2021;9:135550-135558. doi: 10.1109/ACCESS.2021.3117117.

[21] Lebedev O. Scalar overproduction in standard cosmology and predictivity of non-thermal dark matter. J CosmolAstropart Phys. 2023;2023(02):032.

[22] Matin MAA, Triayudi A, Aldisa RT. Comparison of Principal Component Analysis and Recursive Feature Elimination with Cross-Validation Feature Selection Algorithms for Customer Churn Prediction. In: Proceeding of the 3rd International Conference on Electronics, Biomedical Engineering, and Health Informatics: ICEBEHI 2022, 5–6 October, Surabaya, Indonesia. Singapore: Springer Nature Singapore; 2023. p. 203-218.

MULTI DATASET BASED LICENSE PLATE RECOGNITION USING YOLO DETECTORS

*Deepika. P**

ABSTRACT

Detecting the licence plate is the most reliable and economical technique for identifying cars. The methodologies and approaches vary based on several parameters, including but not limited to: picture quality, the vehicle in fixed places, lighting circumstances, and single images. The variations in licence plates across different countries and states must to be able to handle it as well. Additionally, the technique must be able to function properly when there are many characters in the collected photos that have different plate sizes. This endeavor's aim is developing and creating a software for Licence Plate Recognition (LPR), which may be used for e-challan, in car parking, and vehicle identification for smart garage applications through a design by thinking approach. The main focus will be recognising and identifying several automobiles with licence plates from a single picture. Two processes make up the suggested system: plate number identification and identification. The plate number identification procedure identifies the number plate from the photograph, and the segmented plate is then sent to the plate recognition phase in the second phase to determine the characters and numbers.

Keywords: First Keyword, Second Keyword, Third Keyword.

I. INTRODUCTION

Using the licence plate, an image processing technology called licence plate recognitionmay identify the vehicle. The goal is to create an efficient, authorized automated vehicle identification system using the licence plate. The system is installed to monitor safety at the point of entrance into places that are very restricted, including war zones or the vicinity of important government offices, like the Supreme Court or Parliament. This is a system that anybody may use for security reasons. Your phone will be installed with an open-access Android application. After that, all he has to do to get the information he needs on any automobile is to capture a picture of the licence plate and analyse it to extract the required data. This system must be implemented since it is very important. Prior to taking a photograph of the car, the designed equipment detects the vehicle for defensive reasons. The area of the car's number plate is extracted by segmenting the photo into an image. The Optical Character Recognition (OCR) process is used to recognise characters. The obtained information can be then cross-referenced with records to provide unique facts such as location, licencing location, and the owner's name. The structure is developed and emulated in Python then verified on actual picture results.

Simple Licence Plate Recognition (LPR) systems have poor detection accuracy in real-world applications [1]. The accuracy of these systems has been impacted by a number of external factors, including sunlight, headlights from passing cars, number plates with inappropriate designs, and a large variety of number plates. On the other hand, the software and hardware associated with the camera are of limited quality. However, LPR systems are now considerably safer and more widely used due to recent developments in hardware and software [2,3]. These systems are being used by an enormous number of people worldwide, it is expanding rapidly, and it is capable of performing an increasing number of jobs autonomously across various market niches. A result-dependent side programme may correct faults and can provide an almost per- fect system, even if the recognition rate is not 100%. For example, this side programme may disregard certain ignorable faults in the two recognitions to compute the car's parking duration from the time of entering to exiting the parking lot. By cleverly integrating, LPR's drawbacks may be addressed and; dependable, fully automated systems can be created [4, 5,6].

*Department of Artificial Intelligence and Data Science
Karpagam Academy of Higher Education, Coimbatore, Tamil Nadu, Inida
* Corresponding Author