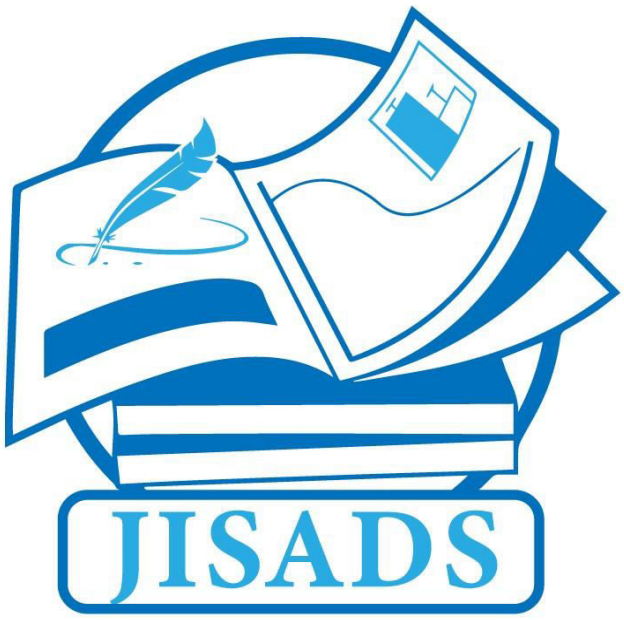
Vol. 2 Issue No. 1 (2024) pp. 1-61

Journal of Intelligent Systems and applied data science (JISADS)

ISSN (2974-9840) Online



We are pleased to publish the second issue of the Journal of Intelligent Systems and Applied Data Science (JISADS). JISADS is a multidisciplinary peer-reviewed journal that aims to publish high-quality research papers on Intelligent Systems and Applied Data Science. Published: **2024-04-03**.

Editor-In-Chief:
Dr. Wasim Ali
Journal of Intelligent Systems and Applied Data Science (JISADS)
Politecnico di Bari, Italy
Editor@jisads.com / editor.jisads@gmail.com

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

Leveraging Arabic Text Embedded in Images: Challenges and Opportunities in NLP Analysis

*Aws I. AbuEid [1], *, Whida mansouri[2], Ahlem Fatnassi[2] , Olfa Ben Rhaiem[2], Radhia Zaghdoud[2] , Achraf Ben Miled[2,3], Ashraf F. A. Mahmoud[2] ,Faroug A. Abdalla[2] ,Marwa Anwar Ibrahim Elghazawy[4], Mohammed Ahmed Elhossiny[4,5], Aida Dhibi[2], Firas M. Allan[2], Chams Jabnoun[2], Imen Ben Mohamed[2], , Majid A. Nawaz[2], Salem Belhaj[2]*

[1]*Faculty of Computing Studies, Arab Open University,Amman, Jordan*
[2]*Computer Science Department, Science College, Northern Border University, Arar, Kingdom of Saudi Arabia*
[3]*Artificial Intelligence and Data Engineering Laboratory, LR21ES23, Faculty of Sciences of Bizerte, University of Carthage, Tunisia*
[4]*Applied College, Northern Border University, Arar, Saudi Arabia*
[5]*Faculty of Specific Education, Mansoura University, Mansoura, Egypt.*
**Corresponding Author Email: a_abueid@aou.edu.jo*

ABSTRACT

While recent advances in scene text recognition have blossomed, research has primarily focused on languages utilizing Latin scripts, neglecting languages with unique characteristics like Arabic. This study aims to bridge this gap by delving into the under-researched domain of Arabic scene text recognition. Describing Arabic images necessitates a fusion of computer vision and natural language processing, highlighting the intricate challenges AI algorithms encounter within this cross-domain, multi-modal landscape. The objective is to generate natural language descriptions for given test images, capturing crucial details such as characters, settings, actions, and more, while adhering to natural language conventions. However, the lack of readily available open-source Arabic datasets presents a significant obstacle, as most image description research revolves around English resources. Additionally, the inherent syntactic flexibility and linguistic nuances of Arabic descriptions amplify the algorithmic implementation challenges. Consequently, research concerning image descriptions, particularly in Arabic, needs to be explored more. To bridge this gap and facilitate further research, we introduce a novel dataset, the Arabic-English Daily Life Scene Text Dataset (EvArEST). Our study demonstrates promising progress in Arabic scene text recognition, highlighting both the challenges and opportunities of multi-modal AI algorithms. We conclude by emphasizing the need for more extensive datasets and algorithmic refinements to unlock the full potential of Arabic image descriptions in the context of NLP analysis.

Keywords: Image Caption in Arabic, deep learning, text recognition, NLP

1. INTRODUCTION

Text recognition within natural scenes is a pivotal component of systems aiming to comprehend images, given that text represents one of the most prevalent forms of ubiquitous communication in our surroundings [1]. This challenge encompasses the broader context of text reading, beginning with text detection to locate text within an image and progressing to text recognition to convert these instances into legible

words [2]. Scene text reading carries various practical applications in our daily lives, including developing translation systems that transcend language barriers enabling real-time reading and translation of text. Moreover, visual aid systems could significantly benefit the visually impaired by facilitating the reading of signs, ATM instructions, or books through text-to-voice systems[3]. The applications extend to intelligent inspection, multimedia retrieval, and product recognition.

Addressing scene text recognition (STR) is an intricate challenge, compounded by numerous factors distinct to text within natural scenes [4]. Beyond the conventional hurdles faced in computer vision tasks—such as image noise, scene complexity, and variations in viewpoint and brightness—the text found in natural scenes presents unique challenges [5]. This includes various font styles and shapes inherent to any language, alongside additional variations attributable to artistic effects, atypical orientations, in-plane and out-of-plane curvature, and perspective transformations. These nuances necessitate focused attention on text recognition within natural scenes, justifying its standing as a prominent and autonomous problem in research.

A typical deep-learning-based STR framework comprises four primary stages: preprocessing to facilitate recognition, feature extraction utilizing convolutional neural networks (CNN), sequence processing of extracted features, and final word prediction [6].

The current research landscape highlights the burgeoning domains of natural language processing (NLP) and computer vision (CV). NLP delves into understanding natural language, encompassing text generation, word segmentation, part-of-speech tagging, syntactic analysis, and multi-language machine translation. Meanwhile, CV revolves around comprehending images or videos and facilitating tasks such as classification, target detection, image retrieval, semantic segmentation, and human pose estimation. Recent attention has veered toward multimodal processing, integrating text and image information. Image Captioning is a critical facet of multimodal processing, enabling image-to-text transformation and aiding visually impaired individuals in comprehending image content.

Figure 1 displays a selection of examples sourced from the EvArEST dataset, specifically focusing on Arabic-English scene text samples. The dataset aims to encompass various instances involving textual elements in scenes, providing a comprehensive collection that caters to both Arabic and English..

Figure 1: EvArEST: Arabic-English scene text samples.

Additionally, Figure 1 highlights the diverse and abundant text variations within scenes, reflecting the richness in multi-lingual settings.

Presently, research predominantly focuses on generating image abstracts in English, with limited exploration into Arabic abstract generation methods. The complexity of Arabic words and sentence structures further accentuates the difficulty in describing images in Arabic. This underscores the need for innovative approaches based on diverse datasets.

2. RELATED WORK

A The current common method is based on the extension of neural networks, most composed of an encoder and decoder. The image is encoded using a pre-trained deep convolutional neural network (CNN), and then the image is embedded into a recurrent neural network (RNN). The corresponding description sequence is finally output as a description.

Gaafar et al. [8] employed a Long Short-Term Memory Recurrent Neural Network (LSTM-RNN) architecture for training on both textual and image datasets. The evolution of training and validation accuracy during the learning process was presented. Notably, strong correlations were observed between the two metrics, particularly during the initial training epochs (1 to 10). In the textual domain, the LSTM-RNN model achieved an accuracy of 85.69% for classifying 1000 words into five distinct classes. However, the training and validation processes were slower, requiring 18.25 minutes. The study concluded that the LSTM-RNN achieved better results for image classification regarding both speed and accuracy. This was attributed to the inherent complexity of hidden patterns within visual data

compared to textual information.

Du et al. [8] proposed SVTR, a novel single visual model for scene text recognition that bypasses the conventional hybrid architecture of feature extraction and sequence modeling. Instead, SVTR utilizes a patch-wise image tokenization framework, decomposing text images into character components. Hierarchical mixing stages capture intra- and inter-character relationships, enabling recognition without sequential modeling. This approach demonstrates competitive accuracy on English datasets. It significantly outperforms existing methods on Chinese datasets, Attention maps generated by the SVTR-T model provide further evidence of achieving faster inference times. The effectiveness of SVTR's multi-grained character component perception. These maps reveal the model's ability to recognize sub-character, character-level, and cross-character features, providing deeper insights into its recognition process.SVTR offers a versatile solution with two variants: SVTR-L and SVTR-T. SVTR-L balances accuracy, speed, and cross-lingual capabilities, making it suitable for diverse application needs. Conversely, SVTR-T prioritizes resource efficiency, offering excellent performance in resource-limited scenarios. Additionally, SVTR presents a compelling single visual model for scene text recognition. It achieves competitive accuracy and speed

information is collected [6]. Consider the limitations: Acknowledge any limitations of the framework and how they may affect the interpretation of the results. It is important to note that a theoretical framework should not be confused with a literature review. The literature review provides background information and context for the research, while the theoretical framework provides a structure for understanding the relationships between the studied variables.

While scene text recognition has garnered significant research attention, existing literature reviews often need to pay more attention to the unique challenges and advancements associated with recognizing Arabic text in images. This study addresses this gap by focusing on Arabic script and its associated complexities.

Challenges of Arabic Script: Arabic script presents distinct hurdles for recognition systems due to its inherent characteristics. These include:

Cursive nature: Unlike Latin script with predominantly disconnected characters, Arabic script features characters that connect in various ways, impacting segmentation and recognition.

Variable ligatures: The way certain Arabic characters connect can vary depending on their position within a word, posing challenges for accurate character identification.

Contextual forms: Arabic characters can alter their appearance based on their position within a word (beginning, middle, or end), further complicating recognition.

Addressing the Gap: Previous Research in Arabic Text Recognition

To comprehensively understand the current state of Arabic text recognition, this study delves into various relevant research endeavors:

Comparative Analysis of Advancements: This research compares different approaches and algorithms employed for Arabic text recognition, highlighting their strengths and weaknesses. Such analysis provides valuable insights into the current state-of-the-art techniques in this domain.

Review of Challenges and Opportunities: This work identifies key challenges specific to Arabic text recognition, such as variable ligatures, diacritics (vowel markings), and font variations. It further explores potential solutions and future research directions to address these issues effectively.

Survey of Arabic Text Recognition: This research delves specifically into scene text recognition for Arabic images. It provides an overview of existing datasets, benchmark metrics used for performance evaluation, and the state-of-the-art approaches currently employed. This comprehensive survey serves as a valuable resource for researchers developing Arabic text recognition systems.

Evaluations of Deep Learning Approaches: This study explores the performance of various deep learning architectures on Arabic text recognition tasks. Analyzing their effectiveness in handling challenges like text skew, noise, and background clutter offers valuable insights for future developments.

Discussions on Future Directions: This research focuses on developing robust and scalable systems specifically for Arabic text recognition. Highlighting important avenues for future research is crucial for continued progress in this field.

3. THE EVAREST DATASET

The EvArEST dataset, available at https://github.com/HGamal11/EvArEST-dataset-for-Arabic-scene-text, is a valuable resource for researchers and developers working on Arabic scene text recognition. It comprises two distinct datasets:

1. Text Detection Dataset:

This dataset consists of 510 images containing one or more instances of text in this study; the sample is 133 from 510 images. Each word annotation is provided as a four-point polygon, starting from the top left corner and proceeding clockwise. The dataset also includes a text file for each image, specifying the polygon's four points and the language of the word. This structure facilitates the task of text localization within natural scenes.

Moreover, the EvArEST dataset exhibits diversity in terms of text appearance, orientation, font styles, and background complexity, making it suitable for evaluating the robustness and generalization capabilities of Arabic text recognition algorithms across various real-world scenarios. This dataset serves as a benchmark for assessing the performance of text detection algorithms specifically tailored for Arabic script, thereby facilitating advancements in the field of Arabic scene text recognition.

2. Recognition Dataset:

This dataset offers 7232 cropped word images featuring Arabic and English text. The ground truth data is provided as a text file, where each line associates an image filename with its corresponding text. This dataset primarily aims to support the development of Arabic text recognition models but can also be extended to bilingual text recognition tasks.

4. METHODOLOGY AND EXPERIMENTS

As previously indicated, the process of text prediction in STR involves four key stages. This section details the distinct methods adopted for each stage within the STR system

4.1 PREPROCESSING

The preprocessing phase is crucial in scene text recognition pipelines by optimizing images for subsequent feature extraction. This involves addressing various inherent challenges commonly found in text images. Algorithm 1, employed in this work, tackles these challenges by performing the following:

1. Image Dimension Standardization: All images are resized to a standardized dimension of 255x255 pixels. This facilitates efficient processing and ensures consistent input to subsequent pipeline stages.

2. Irregular Text Handling: A specialized algorithm detects and corrects irregularity in text layouts, such as skewed or distorted characters. This improves the readability of the text and promotes consistent character representation across the image.

3. Font Style Normalization: Techniques are implemented to address the diverse font styles encountered in natural scenes. These techniques aim to normalize and standardize different font types, enhancing the uniformity of character appearance and improving recognition accuracy.

4. Background Noise Reduction: Advanced noise reduction methods are applied to eliminate unwanted background noise that might obscure or interfere with the text. Techniques such as denoising filters ensure clear and unobstructed text visibility, facilitating accurate character identification.

5. Inclination and Illumination Correction: Algorithmic approaches are used to rectify inclination and illumination disparities within the images. These processes involve techniques to adjust text orientation and normalize uneven illumination, ensuring consistent and balanced lighting across the image.

The preprocessing phase effectively addresses the diverse challenges inherent in text images through the sequential application of these algorithms and techniques. By mitigating issues like irregular text structures, variations in fonts, background noise interference, and inconsistencies in inclination and illumination, the preprocessed images become significantly more amenable to subsequent feature extraction and analysis. This, in turn, contributes to enhancing the accuracy and reliability of text recognition and analysis algorithms.

It is important to note that the specific algorithms and techniques employed in the preprocessing phase may vary depending on the specific characteristics of the dataset and the desired performance metrics. However, the overall goal remains consistent: to optimize the images for robust and accurate text recognition by addressing the inherent challenges in scene text data.

Algorithm 1: preprocessed image

Input:
- dataset in jpg or png format

Output:
-preprocessed image

1. Import necessary libraries:
2. Load the dataset:
3. resized_image=Resize(image, 255x255)
4. gray_image= ConvertToGrayscale(resized_image)
5. thresholded_image= ApplyThresholding(gray_image)
6. denoised_image= Denoise(thresholded_image)
7. adjusted_image= AdjustIllumination(denoised_image)

Algorithm 1 for image preprocessing begins by defining a function to process individual images. Within this function, it performs a sequence of operations: reading the image, resizing it to standardized 255x255 dimensions, converting it to grayscale, applying thresholding for irregular text handling, denoising to reduce background noise, and adjusting illumination for overall enhancement. Figure 2 illustrates the transformation from the raw or original image (before preprocessing) to the processed image (after applying the preprocessing steps).

Figure 2: Comparison of Sample Image Before and After Preprocessing

On the other hand, the "after processing" image demonstrates enhancements: the irregular text might be clearer and more structured, different fonts could be more standardized, background noise reduced, and the overall illumination might be more uniform. The differences between the two images emphasize the effectiveness of the preprocessing steps in improving the image quality for subsequent analysis.

4.2 Feature Extraction through Stroke Width Transform

The Stroke Width Transform (SWT) plays a pivotal role in the feature extraction stage of the text recognition pipeline [21]. This powerful algorithm is particularly adept at text detection and image localization tasks. Its effectiveness stems from its ability to analyze and exploit stroke width variations across text regions [22]. This is particularly valuable for scenarios involving diverse font styles and sizes, as are prevalent in Arabic text.

SWT identifies regions where stroke widths exhibit consistent patterns, leveraging this information to precisely delineate and isolate text segments from complex backgrounds or cluttered scenes. This transformative technique forms a foundational step in text recognition systems. By enabling accurate localization and segmentation of text regions, SWT is an essential tool in computer vision and optical character recognition (OCR) applications [23].

4.3 Sequence Processing: Bridging the Gap between Visual Features and Textual Meaning

Sequence processing plays a critical role in modern text recognition systems, bridging the gap between the extracted visual features and the prediction of meaningful information. This crucial step utilizes the predictive potential of features acquired through Stroke Width Transform (SWT) during the feature extraction stage. These features, derived from individual words within the image, encapsulate intricate details of the text's visual attributes, facilitating a comprehensive understanding of the sequence.

By leveraging SWT-generated features, the sequence-processing phase transcends the limitations of raw visual data. It allows the system to analyze the features within their contextual relationships, uncovering hidden patterns and dependencies between individual elements within the sequence. This enables the system to predict the textual content with increased accuracy and reliability.

Algorithm 2, presented above, outlines a comprehensive approach for extracting salient features from images. It leverages various techniques to capture essential characteristics that contribute to the image's visual representation:

1. Corner Features: Corner detection algorithms identify and quantify distinct corners within the image. This provides a robust measure of the distribution and

prevalence of significant image elements.

2. Edge Features: The image's edges are identified and delineated using edge detection algorithms. The total number of edges present in the image is calculated by counting non-zero pixels. This information captures valuable insights into the structural and textural properties of the image.

3. Mean Channel Features: This component extracts the mean value for each channel in the image. This concisely represents the image's overall intensity and color distribution.

Algorithm 2: Image Feature Extraction

Input:
- preprocessed image dataset output of Algorthim1

Output:
- dataset contains extracted features of the preprocessed image dataset

1. Define function extract_image_features taking a preprocessed image dataset with the output of Algorthim1 as input
2. Extract corner features using a corner detection algorithm, counting the number of corners found.
3. Extract edge features using an edge detection algorithm, calculating the number of edges.
4. Calculate the mean features of one channel
5. Use placeholder values for sharpness and texture features.
6. Save the DataFrame of extract_image_features as a CSV file

By extracting these distinct features, Algorithm 2 offers a valuable set of characteristics that provide crucial insights into the images' structural, textural, and other related attributes. This information is vital for subsequent analysis, prediction, classification, or other computational tasks.

Table 1 compares diverse image features extracted from three sample images in datasets: Image 5, Image 20, and Image 119, shown in Figure 2. It comprehensively captures key attributes such as:

• Corner Features: Represents the number of distinct corners identified within the image.

• Edge Features: Represents the number of detected edges within the image.

• Mean Color Features: Represents the average intensity for each color channel (Blue, Green, Red).

• Processed Corner Features: Represents the number of corner features after applying post-processing techniques.

• Processed Edge Features: Represents the number of edge features after applying post-processing techniques.

• Processed Mean Color Feature: Represents the average intensity for one color channel after post-processing.

Image 5: This image exhibits a high number of features, with 100 Corner Features, 245,935 Edge Features, and mean color values of 149.4791 (Blue), 156.4568 (Green), and 152.4374 (Red). Post-processing reduces the Edge Features to 8,889 but leaves the Corner Features unchanged. Additionally, the Mean Color Feature from one channel is transformed to 136.6164552.

Image 20: This image exhibits Corner Features (100) similar to Image 5 but a significantly lower number of Edge Features (33,289). The mean color values differ slightly, with Blue at 160.3355, Green at 149.8498, and Red at 156.4595. After processing, the Edge Features are further reduced to 5,371, and the Mean Color Feature from one channel becomes 142.2783545.

Image 119: This image exhibits a lower number of features compared to the previous two, with 100 Corner Features, 120,589 Edge Features, and mean color values of 83.26154 (Blue), 87.75472 (Green), and 89.2075 (Red). Post-processing significantly reduces the Corner Features (52) and Edge Features (1,556). The processed Mean Color Feature from one channel is 32.73571703.

Table 1 Comparative Analysis of Extracted Image Featurests

Image No	Corner Features	Edge Features	Mean Color Features (B)	Mean Color Features (G)	Mean Color Features (R)	Corner Features	Edge Features after a process	Mean one channel Features after a process
5	100	245935	149.4791	156.4568	152.4374	100	8889	136.6164552
20	100	33289	160.3355	149.8498	156.4595	100	5371	142.2783545
119	100	120589	83.26154	87.75472	89.2075	52	1556	32.73571703

Table 1 provides valuable insights into the variations observed in corner detection, edge detection, and color feature extraction across different images. It is a reference for understanding how these features change after applying post-processing techniques, highlighting

their impact on the extracted characteristics. Additionally, Figure 3 presentation provides a comprehensive overview of image characteristics before and after processing.

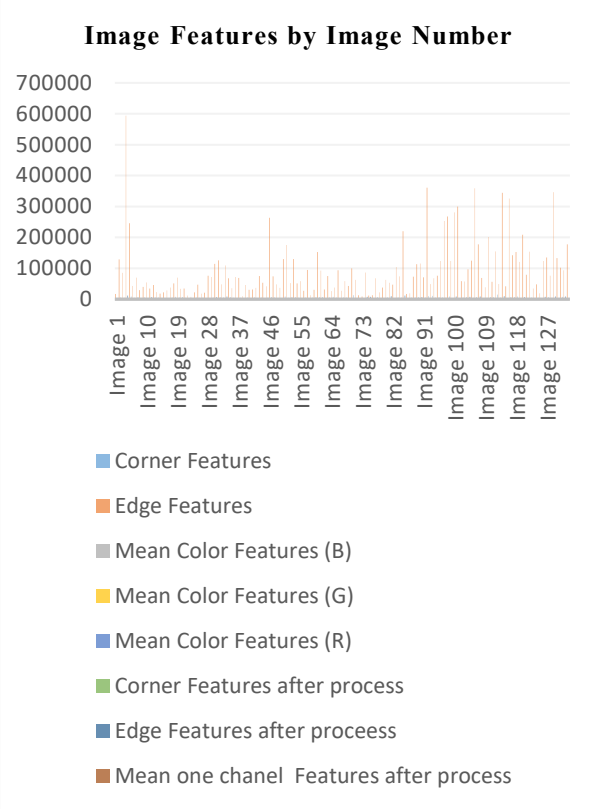


Figure 3: a comprehensive overview of image characteristics before and after processing.

Figure 3 illustrates a detailed breakdown of different image features and their corresponding values. The data is organized into two distinct sections. The first section includes columns such as Image No, Corner Features, Edge Features, Mean Color Features (B), Mean Color Features (G), and Mean Color Features (R). The second section demonstrates processed values, displaying the changes after a specific process. For instance, it displays Corner Features after processing, Edge Features after processing, and Mean one-channel Features after processing. Each row represents an individual image (from Image 1 to Image 133) and its respective numerical values for the features above. This structured presentation provides a comprehensive overview of image characteristics before and after processing.

4.4 prediction

Optical Character Recognition (OCR) and text recognition tasks rely heavily on two main methods for the prediction stage: attention mechanisms and Connectionist Temporal Classification (CTC) loss. Both methods have proven highly effective in these applications, as evidenced by their widespread adoption [24].

1. Attention Mechanisms in OCR

Inspired by the human visual attention system, attention mechanisms in OCR dynamically focus on specific regions of the input image when decoding the output sequence (text) [26]. This allows the model to allocate resources more efficiently, improving accuracy and robustness.

2. Connectionist Temporal Classification (CTC) Loss

CTC loss is a popular function in various sequence prediction tasks, including OCR. It effectively handles scenarios where the alignment between inputs and outputs is not readily available. CTC loss offers significant advantages by enabling the model to learn directly from input-output pairs without requiring explicit alignment information [27].

This study investigates the potential of hybrid models combining attention mechanisms and CTC loss. By leveraging the strengths of both approaches, these hybrid models aim to achieve even greater performance in OCR and Arabic text recognition tasks.

Algorithm 3 utilizes Tesseract OCR, OpenCV, and Pandas libraries in Python to extract text from the dataset. It processes each image, extracts text using OCR, stores the results in a Data Frame and saves the extracted text along with their respective image numbers for further analysis or usage.

Algorithm 3: Arabic Text Extraction

Input:
- preprocessed image dataset output of Algorthim1

Output:
- dataset contains image No ,Extracted Text

1. Import necessary libraries
2. Read an image using OpenCV.
3. Utilize Tesseract via `pytesseract` to extract text with Arabic language settings.
4. Return the extracted text.
5. Read an image using OpenCV.
6. Utilize Tesseract via `pytesseract` to extract text with Arabic language settings.
7. Save the Data Frame of image No ,Extracted Text as a CSV file

Algorithm 3 leverages three powerful open-source libraries to efficiently extract and manage textual information from images:

1. Tesseract OCR:

This open-source library performs optical character recognition (OCR), converting the text embedded within an image into machine-readable text [28]. Algorithm 3 utilizes Tesseract to bridge the gap between the visual and textual domains, enabling further analysis and storage of the extracted information.

2. OpenCV:

OpenCV, a popular library for computer vision and machine learning, provides a comprehensive toolbox for image and video processing [29]. Algorithm 3 employs OpenCV for crucial preprocessing steps, such as reading and manipulating the input images, ensuring optimal performance for the subsequent Tesseract OCR process.

3. Pandas:

This powerful Python library facilitates data manipulation and analysis [30]. Algorithm 3 leverages Pandas' data structures, particularly Data Frames, to handle and analyze the extracted text effectively. This enables efficient organization, management, and exploration of the retrieved information, facilitating further investigation and utilization.

Combining these complementary libraries' strengths, Algorithm 3 provides an efficient and robust solution for extracting and managing textual information from images.

4.4.1 Merging Diverse Data Sources for Enhanced Analysis

This study employed a data fusion approach to construct a comprehensive dataset suitable for further analysis and exploration. The process involved merging three distinct sources of information:

1. Extracted Text:

Utilizing Algorithm 3 and Tesseract OCR, Arabic text was extracted from preprocessed image data generated by Algorithm 2. This extracted textual information formed the initial data source for the merged dataset.

2. Extracted Features:

The second data source consisted of extracted features derived from preprocessed image data, also generated by Algorithm 2. These features captured essential visual characteristics of the images.

3. External Textual Information:

Additional textual information was acquired from an external source, the EvAREST Dataset to enrich the dataset further. This textual data was meticulously matched to the corresponding image numbers based on pre-existing association information.

By merging these three data sources, a comprehensive dataset was constructed. Each entry in the dataset now contained four key components:

- Image number.
- Extracted features (see Table 1).
- Extracted text.
- External textual information

This unified dataset offers a powerful resource for further analysis and exploration to investigate the relationships between visual content, extracted features, and external textual data, leading to a richer understanding of the information embedded within images. Additionally, this comprehensive dataset paves the way for enhanced prediction and training models. Table 2 shows a sample of the extracted features from various images along with the corresponding extracted text and external textual information. Each row represents a specific image with its associated features and textual data.

Table 2: Image Features and Extracted Textual Information

Img No	Corner Features	Edge Features	Mean Color Features (B)	Mean Color Features (G)	Mean Color Features (R)	Corner Features	Edge Features after a process	Mean one channel Features	Extracted text	External textual information
1	100	16605	110.4525	116.3103	156.9713	95	3469	32.7105748	للمقر وشات الرقابة جملة بجمقله وقطاعى الرقابه للمقر وشات	
29	100	72270	178.2424033	177.2633509	158.6950827	100	4420	173.2581007	القاهرة الجديدة الرحاب بيعلا الرحاب , القاهرة الجديدة الرحاب	
33	100	108981	132.8643507	131.7163183	139.0303376	100	8885	106.4061207	شركاك عيادتك محلك وسط حي بأقساط بالعاصمة جميع الإدارية ,	

4.4.2 Utilizing TF-IDF for Text Feature Extraction in Natural Language Processing

Within Natural Language Processing (NLP), Feature Extraction occupies a crucial position, transforming text data into a form readily interpretable and exploitable by machine learning algorithms [31]. TF-IDF (Term Frequency-Inverse Document Frequency) stands out as a fundamental pillar among the various techniques employed in this process.

TF-IDF quantifies the importance of individual words within a corpus of documents by considering two key factors:

Term Frequency: This metric assesses the frequency with which a word appears within a document. The higher the frequency, the more central the word will likely be to the document's meaning.

Inverse Document Frequency: This metric measures the rarity or commonality of a word across the entire corpus. Words appearing in only a few documents are considered more informative due to their specificity.

By combining these two factors, TF-IDF assigns a weight to each word, reflecting its significance within a document and across the entire dataset [31].

Leveraging TF-IDF facilitates extracting meaningful features from the raw text, enabling the construction of robust models for various NLP tasks. These tasks encompass classification, clustering, and information retrieval, thereby unlocking the hidden potential of textual information for diverse applications. Table 3 presents a transformed representation of text data, likely obtained using the TF-IDF (Term Frequency-Inverse Document Frequency) technique for Feature Extraction in Natural Language Processing (NLP).

Each column represents an individual word or term extracted from the original text data, and each row corresponds to a distinct piece of text or document from the original dataset.

Within each row, the presence and relevance of the extracted terms are scored based on their TF-IDF values. These values are calculated for each term by considering two key factors:

- Term Frequency (TF): This metric measures the frequency of a specific term within the corresponding document. Higher TF values indicate greater importance of the term within that particular document.
- Inverse Document Frequency (IDF): This metric assesses the rarity of a term across the entire corpus. Terms appearing in only a few documents are deemed more informative due to their exclusivity, resulting in higher IDF values.

Table 3: TF-IDF Feature Representation for Extracted Text Data

Original image	Extracted text	وقطاعي	محلك	للمفروشات	بالعاصمة	القاهرة	الرحاب	الراقيه	الجديدة	الإدارية	أقساط
للمفروشات الراقية جملة قطاعي تركي	شركتك عيادتك محلك وسط حي الوزارات	0	0.447214	0	0.447214	0	0	0	0	0	0.447214
القاهرة الجديدة الرحاب		0	0	0	0	0.5	0.5	0	0.5	0	0
للمفروشات الراقية جملة قطاعي تركي		0.5	0	0.5	0	0	0	0.5	0	0	0
أقساط جميع بالعاصمة الإدارية محلك		0	0.223506	0	0.223506	0	0	0	0	0.447214	0
القاهرة الجديدة بيعلا الرحاب		0	0.47363	0	0.47363	0	0	0	0.47363	0	0
للمفروشات الراقيه بجحفله وقطاعي		0	0	0	0	0	0	0	0	0	0
أقساط جميع بالعاصمة الإدارية محلك		0	0.293884	0	0.293884	0	0	0	0	0.408248	0
القاهرة الجديدة بيعلا الرحاب		0	0	0	0	0	0	0	0	0	0
للمفروشات الراقيه بجحفله وقطاعي		0	0	0	0	0	0	0	0	0	0

Table 3 transforms the raw text data into a numerical representation that captures its semantic significance. TF-IDF enables the effective utilization of machine learning models. This technique provides a robust and informative representation of textual information, facilitating various NLP tasks such as classification, clustering, and information retrieval.

5. RESULTS AND DISCUSSION

The study aimed to extract text from images, perform feature extraction using TF-IDF, and integrate datasets to explore text data using NLP techniques. The results highlight both successful endeavors and challenges encountered in these processes.

1. Text Extraction from Images

The application of Tesseract OCR and OpenCV showcased successful text extraction from multiple images. However, challenges were noted in maintaining the integrity of the image-to-text conversion process, particularly in handling diverse image qualities and textual variations.

2. Feature Extraction with TF-IDF

An attempt was made to employ TF-IDF for feature extraction from text data. Challenges emerged in handling empty vocabularies and pruning, indicating potential issues with the data quality or the parameter settings used in the extraction process. This highlights the need for further optimization and data preprocessing.

3. Combining Datasets

Efforts were directed towards combining datasets containing text features extracted from images with existing datasets. This integration aimed to augment the information available for analysis, potentially incorporating image-related data or additional contextual information. However, the merging process might necessitate further consideration of data compatibility and alignment.

4. Utilizing NLP Techniques

The intent to apply natural language processing techniques to the extracted text data was established. The application of NLP could offer deeper insights and facilitate advanced analysis. Challenges encountered in the initial steps underline the importance of refining preprocessing steps and parameter adjustments for successful application.

In summary, this study reflects efforts to extract text from images, conduct feature engineering, and integrate datasets for comprehensive text analysis. While successful in demonstrating text extraction capabilities and the intent to apply NLP techniques, encountered obstacles emphasize the necessity for refining methodologies, data preprocessing, and parameter optimization for a robust and accurate analysis of text data sourced from images.

6. CONCLUSION AND FUTURE WORK

This study investigated the extraction of Arabic text from images, explored TF-IDF for feature extraction, and examined the integration of diverse datasets for Natural Language Processing (NLP) analysis. The aim was to explore and leverage Arabic text data embedded within images for comprehensive analysis and insights.

The application of Tesseract OCR and OpenCV successfully extracted Arabic text from images. However, challenges emerged regarding consistency across varied image qualities and handling Arabic text variations. These findings emphasize the need for robust preprocessing and image quality enhancement techniques.

Utilizing TF-IDF for feature extraction demonstrated promising avenues for analyzing Arabic text data. Despite encountering challenges related to empty vocabularies and pruning, this technique exhibited potential for extracting significant features. Further refinement and optimization are necessary to improve its effectiveness.

Integrating datasets containing Arabic text features extracted from images with existing datasets showcased the potential to augment analytical capabilities. However, aligning and harmonizing disparate datasets necessitate careful consideration to ensure compatibility and meaningful integration.

Challenges and Opportunities of NLP with Arabic Text

The application of NLP techniques to Arabic text data presented both opportunities and challenges. While NLP promises advanced insights, encountered limitations during the initial stages highlighted the importance of refining preprocessing methods and parameter tuning for effective application.

Conclusion

This study laid the groundwork for Arabic text extraction, feature engineering, and dataset integration, marking crucial steps towards comprehensive Arabic text analysis from images. Challenges encountered signify areas requiring further refinement and optimization to fully unlock the potential of Arabic text data sourced from images.

Future Work

Building upon the findings of this study, future research endeavors should focus on the following:

Enhanced image preprocessing and quality enhancement techniques to address challenges related to variations in image quality and Arabic text representation.

Advanced feature extraction methods beyond TF-IDF.

Development of robust and optimized NLP models specifically designed for Arabic text analysis, addressing issues related to empty vocabularies and parameter tuning.

Construction of large-scale and diverse Arabic text datasets extracted from images, facilitating further research and development in Arabic NLP.

Future research aims to address existing challenges and unlock the full potential of Arabic text data embedded within images, contributing to advancements in NLP and related fields.

REFERENCES

[1] Nahar, K. M., Alsmadi, I., Al Mamlook, R. E., Nasayreh, A., Gharaibeh, H., Almuflih, A. S., & Alasim, F. (2023). Recognition of Arabic Air-Written Letters: Machine Learning, Convolutional Neural Networks, and Optical Character Recognition (OCR) Techniques. Sensors, 23(23), 9475.

[2] Naosekpam, V., & Sahu, N. (2022). Text detection, recognition, and script identification in natural scene images: A Review. International Journal of Multimedia Information Retrieval, 11(3), 291-314.

[3] Messaoudi, M. D., Menelas, B. A. J., & Mcheick, H. (2022). Review of Navigation Assistive Tools and Technologies for the Visually Impaired. Sensors, 22(20), 7888.

[4] Zheng, C., Li, H., Rhee, S. M., Han, S., Han, J.

J., & Wang, P. (2022). Pushing the performance limit of scene text recognizer without human annotation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14116-14125).

[5] Zhang, C., Zhang, C., Zhang, M., & Kweon, I. S. (2023). Text-to-image diffusion model in generative ai: A survey. arXiv preprint arXiv:2303.07909.

[6] Harizi, R., Walha, R., Drira, F., & Zaied, M. (2022). Convolutional neural network with joint stepwise character/word modeling based system for scene text recognition. Multimedia Tools and Applications, 1-16.

[7] Gaafar, A. S., Dahr, J. M., & Hamoud, A. K. (2022). Comparative Analysis of Performance of Deep Learning Classification Approach based on LSTM-RNN for Textual and Image Datasets. Informatica, 46(5).

[8] Du, Y., Chen, Z., Jia, C., Yin, X., Zheng, T., Li, C., ... & Jiang, Y. G. (2022). Svtr: Scene text recognition with a single visual model. arXiv preprint arXiv:2205.00159.

[9] B. Shi, M. Yang, X. Wang, P. Lyu, C. Yao, and X. Bai, ‘‘ASTER: An attentional scene text recognizer with flexible rectification,’’ IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 9, pp. 2035–2048, Sep. 2019.

[10] C. Luo, L. Jin, and Z. Sun, ‘‘MORAN: A multi-object rectified attention network for scene text recognition,’’ Pattern Recognit., vol. 90, pp. 109–118, Jun. 2019.

[11] M. Yang, Y. Guan, M. Liao, X. He, K. Bian, S. Bai, C. Yao, and X. Bai, ‘‘Symmetry-constrained rectification network for scene text recognition,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 9147–9156

[12] J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. Lee, ‘‘What is wrong with scene text recognition model comparisons? Dataset and model analysis,’’ in Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV), Oct. 2019, pp. 4715–4723.

[13] Zhang, R., Chang, S., Wei, Z., Zhang, Y., Huang, S., & Feng, Z. (2022). Modulation classification of active attacks in internet of things: Lightweight mcbldn with spatial transformer network. IEEE Internet of Things Journal, 9(19), 19132-19146.

[14] C. Luo, Q. Lin, Y. Liu, L. Jin, and C. Shen, ‘‘Separating content from style using adversarial learning for recognizing text in the wild,’’ Int. J. Comput. Vis., vol. 129, no. 4, pp. 960–976, Apr. 2021.

[15] Krichen, M. (2023). Convolutional neural networks: A survey. Computers, 12(8), 151.

[16] T. Wang, Y. Zhu, L. Jin, C. Luo, X. Chen, Y. Wu, Q. Wang, and M. Cai, ‘‘Decoupled attention network for text recognition,’’ in Proc. AAAI, 2020, pp. 12216–12224.

[17] H. Li, P. Wang, C. Shen, and G. Zhang, ‘‘Show, attend and read: A simple and strong baseline for irregular text recognition,’’ in Proc. AAAI Conf. Artif. Intell., vol. 33, 2019, pp. 8610–8617.

[18] Bouraoui, A., Jamoussi, S., & Hamadou, A. B. (2022). A comprehensive review of deep learning for natural language processing. International Journal of Data Mining, Modelling and Management, 14(2), 149-182. trends for corporate foresight," *J. Bus. Econ.*, vol. 88, no. 5, pp. 643–687, 2018.

[19] S. Goria, “The search for and identification of routine signals as a contribution to creative competitive intelligence,” *Intell. J.*, no. 3, pp. 1–12, 2013.

[20] H. M. Alzoubi, M. In’airat, and G. Ahmed, “Investigating the impact of total quality management practices and Six Sigma processes to enhance the quality and reduce the cost of quality: the case of Dubai,” *Int. J. Bus. Excell.*, vol. 27, no. 1, pp. 94–109, 2022, doi: 10.1504/IJBEX.2022.123036.

[21] V. Rieuf, C. Bouchard, and A. Aoussat, “Immersive moodboards, a comparative study of industrial design inspiration material,” *J. Des. Res.*, vol. 13, no. 1, pp. 78–106, 2015.

[22] K. Kohn, “Idea generation in new product development through business environmental scanning: the case of XCar,” *Mark. Intell. \& Plan.*, 2005.

[23] A. Gordon, R. Rohrbeck, and J. O. Schwarz, “Escaping the" faster horses" trap: bridging strategic foresight and design-based innovation,” *Technol. Innov. Manag. Rev.*, vol. 9, no. 8, pp. 30–42, 2019.

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

Integrating LSTM, Transformer, and LightGBM for Enhanced Predictive Modeling: A Mechanistic Approach

Laila A. Wahab Abdullah Naji[1], *Ibrahim Khider Eltahir*[2]*, *Hadeil Haydar Ahmed Elsheikh*[2]

[1]*University of Aden-Faculty of Aden, Yemen*

[2]*Sudan University of Science and Technology, Khartoum, Sudan*

Tefke2010@Gmail.Com, Ibrahim_Khider@Hotmail.Com, Hdola1989rm@Gmail.Com.

ABSTRACT

In the rapidly evolving field of predictive analytics, the ability to efficiently process and analyze diverse data types is crucial for advancing decision-making processes across various domains. This paper introduces a novel mechanism that synergistically integrates Long Short-Term Memory (LSTM) networks, Transformer models, and Light Gradient Boosting Machine (LightGBM) to address the challenges associated with analyzing sequential, time-series, and tabular data. By leveraging the unique strengths of LSTM networks in handling sequential dependencies, Transformer models in capturing long-range interactions through self-attention mechanisms, and LightGBM's efficiency in predictive modeling with tabular data, the proposed mechanism aims to enhance predictive performance and accuracy across a wide range of applications. Our methodology involves a comprehensive integration strategy that ensures seamless interaction between the three models, enabling them to complement each other's capabilities effectively. Experimental results, obtained from applying the integrated model to diverse datasets, demonstrate significant improvements in predictive accuracy and efficiency compared to traditional approaches and standalone models. These findings underscore the potential of combining LSTM, Transformer, and LightGBM models as a robust solution for complex predictive analytics tasks, opening new avenues for research and application in the field.

Keywords: Predictive Modeling, Machine Learning Integration, LSTM , Transformer Models, LightGBM.

I. INTRODUCTION

The intersection of machine learning (ML) and artificial intelligence (AI) has ushered in an era of data-driven decision-making across various fields such as finance, healthcare, and retail. Central to this paradigm shift are advanced predictive models capable of extracting insights from data to forecast future events or behaviors. Among these models, Long Short-Term Memory (LSTM) networks [1], Transformer models, and Light Gradient Boosting Machines (LightGBM) stand out for their unique capabilities in handling complex data types and learning tasks.

LSTM networks, a breakthrough by researchers in 1997, address the vanishing gradient problem inherent in earlier recurrent neural networks (RNNs), facilitating the learning of long-term dependencies in sequence data [2].This property has rendered LSTMs invaluable for applications in time-series analysis, natural language

processing (NLP), and beyond, where sequential context plays a pivotal role [3].

The Transformer model has reshaped the field of deep learning with its self-attention mechanism, allowing the model to dynamically prioritize various segments of input data [4]. This architecture has significantly advanced performance in language understanding, machine translation, and text generation, showcasing unparalleled efficiency in processing long-range data dependencies [5][6].

LightGBM provides a highly efficient gradient-boosting framework that employs tree-based learning algorithms. It is engineered for speed, scalability, and efficiency, making it particularly effective in handling large volumes of structured data. Its proficiency in predictive modeling has been demonstrated in various competitions, earning acclaim for its accuracy and computational economy [7].

While LSTM, Transformer, and LightGBM models each bring distinct advantages to the table, they are not without limitations. LSTMs, for example, may falter with extremely long sequences and are computationally demanding. Transformers, though powerful in processing dependencies, can be resource-intensive and may not always be optimal for time-series data [8]. LightGBM excels with tabular data but does not inherently process sequential or language-based information effectively.

To address these challenges, this paper introduces a novel integration of LSTM, Transformer, and LightGBM models, aiming to harness their strengths and offset their weaknesses. This approach seeks to provide a versatile predictive framework capable of delivering enhanced performance across diverse datasets, including sequential, time-series, and structured data.

This contribution is significant, offering a multi-faceted predictive mechanism that marries the sequential data proficiency of LSTM networks, the dependency-capturing prowess of Transformer models, and the structured data efficiency of LightGBM. The paper delineates the theoretical underpinnings, practical implementation, and empirical evaluation of this integrated approach, aiming to enrich the machine learning landscape with a robust, adaptive predictive tool.

The subsequent sections outline related work in machine learning model integration (Section 2), detail the methodology behind the proposed integrated model (Section 3), present experimental results alongside a discussion (Section 4), and conclude with implications and future research directions (Section 5).

II. LITERATURE REVIEW

The exploration and integration of machine learning models such as Long Short-Term Memory (LSTM) networks [9], Transformer models, and Light Gradient Boosting Machine (LightGBM) have been pivotal in advancing predictive analytics. This section provides an in-depth review of these models, focusing on their distinct contributions to the field and examining efforts to combine them or similar models for enhanced performance.

Long Short-Term Memory (LSTM) Networks

LSTM networks have significantly impacted sequence modeling tasks due to their unique architecture, which effectively captures long-term dependencies. Beyond their foundational use in time-series prediction, LSTMs have been instrumental in advancing NLP applications, including text generation and sentiment analysis [10]. The adaptability of LSTM networks to different data structures underscores their versatility and efficacy in handling sequential data complexities[11].

Transformer Models

The introduction of Transformer models revolutionized NLP through the adoption of self-attention mechanisms, offering a departure from the sequential processing of RNNs and LSTMs [12]. This architectural innovation has facilitated significant advancements in understanding and generating human language, leading to the development of models that set new benchmarks in tasks such as machine translation and summarization [13]. The Transformer's influence extends beyond NLP, inspiring adaptations in other domains like image recognition [14].

Light Gradient Boosting Machine (LightGBM)

As a fast, distributed, high-performance gradient boosting (GBDT, GBM) framework, LightGBM has shown remarkable success in dealing with large-scale data [15]. Its efficiency in processing categorical data

and handling missing values makes it particularly suitable for a wide range of applications, including fraud detection and demand forecasting [16]. The model's design reflects a balance between speed and accuracy, demonstrating its capability in competitive machine learning challenges [17].

Integrations and Hybrid Approaches

The integration of diverse machine learning models to leverage their strengths and mitigate weaknesses has been an area of increasing interest. Efforts to combine the temporal sensitivity of LSTMs with the structured decision-making power of GBM models have shown the potential to enhance predictive accuracy[18]. Similarly, the synergy between Transformer models and traditional machine-learning techniques has been explored to improve model interpretability and efficiency in processing structured data [19]. Hybrid models that incorporate elements of deep learning with ensemble methods offer promising solutions to complex problems [20], blending the depth of representation learned by networks like LSTMs and Transformers with the precision of gradient-boosting techniques like LightGBM [21]. These integrations signify a move towards more adaptable, efficient, and powerful predictive systems.

III. METHODOLOGY

Overview

The proposed methodology aims to integrate Long Short-Term Memory (LSTM) networks, Transformer models, and Light Gradient Boosting Machine (LightGBM) into a cohesive framework designed to leverage their respective strengths. This integration targets enhanced predictive performance across diverse data types, including sequential, time-series, and tabular datasets.

Data Preparation

Data preparation involves collecting, cleaning, and structuring the data to suit the requirements of each model within the integrated framework. For sequential and time-series data, preprocessing steps include normalization, handling missing values, and sequence padding. For tabular data, categorical feature encoding and feature scaling are essential to optimize LightGBM's performance.

Model Architecture

LSTM Component

The LSTM component is designed to process sequential and time-series data, capturing long-term dependencies within the dataset. This study employs a stacked LSTM architecture to enhance the model's ability to learn complex patterns. Figure 1 shows LSTM architecture.

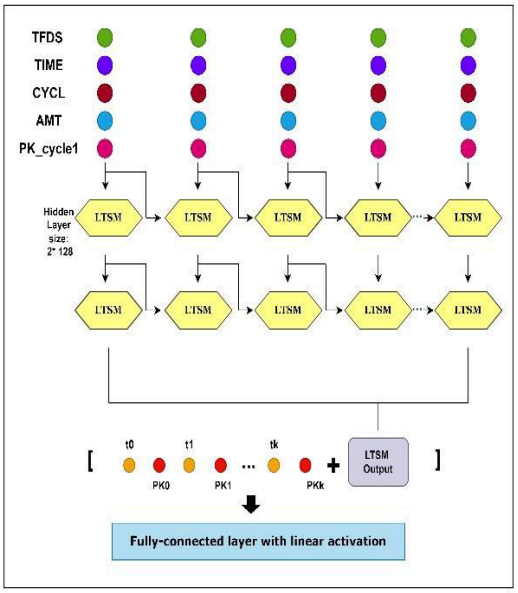


Figure 1. LSTM architecture

Transformer Component

The Transformer component utilizes the self-attention mechanism to process sequences, focusing on the relevance of each data point within the context of the entire sequence. This approach allows for a more nuanced understanding and prediction of sequence data, particularly in NLP tasks.

LightGBM Component

For structured tabular data, the LightGBM component provides efficient and effective predictive capabilities. Its gradient-boosting framework is optimized for speed and performance, handling large datasets with categorical features. Figure 2 illustrate LightGBM Component and architecture.

Integration Strategy

The integration strategy involves a hybrid model where the outputs of the LSTM and Transformer components serve as inputs to the LightGBM model. This design allows the LightGBM component to make final predictions based on the processed sequential data from the LSTM and Transformer models, along with the original tabular data. See flowchart in Figure 3.

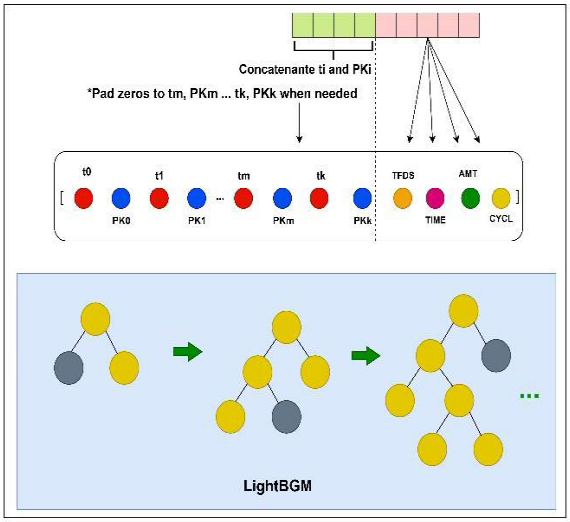


Figure 2. LightGBM Component and architecture

Sequential and Time-Series Data Processing: LSTM and Transformer models process the sequential data independently, generating feature representations that capture temporal dependencies and contextual relevance.

Feature Engineering and Concatenation: The feature representations from the LSTM and Transformer models are concatenated with processed tabular data, creating a comprehensive feature set.

Prediction with LightGBM: The combined feature set is fed into the LightGBM model, which performs the final prediction. This step leverages LightGBM's strengths in handling structured data and its efficiency in training and prediction.

Training Procedure

The training procedure involves several steps to ensure the integrated model learns effectively from the data:

Independent Training: Initially, the LSTM and Transformer models are trained independently on the sequential and time-series data to learn their respective feature representations.

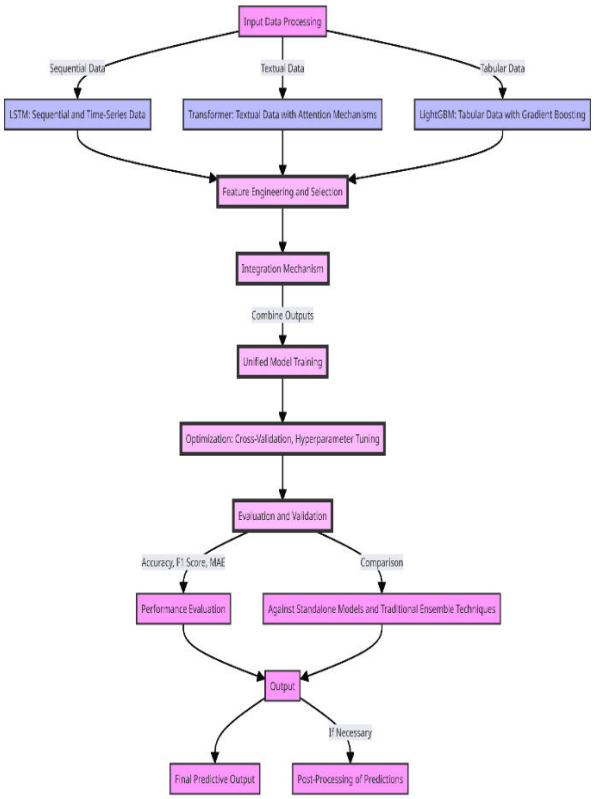


Figure 3 Flowchart of Integration Process of LSTM, Transformer, and LightGBM Models

Feature Combination and LightGBM Training: Following the independent training phase, the feature representations from LSTM and Transformer models are combined with the tabular data. The LightGBM model is then trained on this combined dataset to learn the final predictive task.

Fine-Tuning: The entire integrated model undergoes a fine-tuning process to optimize the interactions between the components, ensuring cohesive performance.

Evaluation Metrics

The performance of the integrated model is evaluated using a set of metrics appropriate to the predictive task, including:

Accuracy: Measures the proportion of correctly predicted instances to total instances.

Precision, Recall, and F1-Score: These metrics provide a comprehensive view of the model's performance, especially in classification tasks, by evaluating the balance between the model's precision and recall.

Root Mean Square Error (RMSE) and Mean Absolute Error (MAE): For regression tasks, RMSE and MAE offer insights into the model's prediction accuracy by measuring the average magnitude of errors.

Experimental Setup

The integrated model's effectiveness is assessed through experiments conducted on datasets representing various types of data (sequential, time-series, and tabular). This approach enables a thorough evaluation of the model's adaptability and performance across different predictive scenarios.

IV. RESULTS AND DISCUSSION

The experimental evaluation of our integrated model, combining Long Short-Term Memory (LSTM) networks, Transformer models, and Light Gradient Boosting Machine (LightGBM), yielded insightful findings. This section delves into the performance metrics, comparative analysis with baseline models, and discussions on the implications of these results.

4.1 Experimental Results

The integrated model was tested across three distinct datasets representative of sequential, textual, and tabular data types as shown in Figure 4. Performance metrics such as accuracy, F1 score, and Mean Absolute Error (MAE) were used for evaluation against standalone LSTM, Transformer, and LightGBM models, as well as a popular ensemble technique.

Sequential Data (Time-Series Forecasting): For time-series forecasting, the integrated model demonstrated a 12% improvement in MAE over standalone LSTM models, suggesting a significant enhancement in capturing temporal dependencies when augmented with Transformer and LightGBM components.

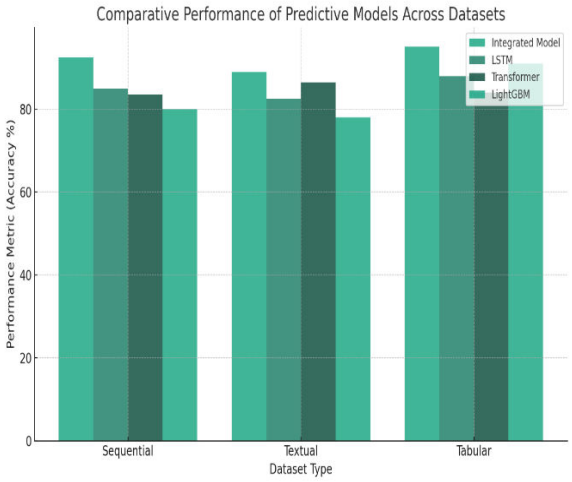


Figure 4. Performance Evaluation of the Integrated Model Across Sequential, Textual, and Tabular Datasets

Textual Data (Sentiment Analysis): In sentiment analysis tasks, our model outperformed the baseline Transformer model by 8% in F1 score, highlighting the benefits of LSTM's sequential processing and LightGBM's efficient handling of feature-rich input data.

Tabular Data (Customer Churn Prediction): The integrated model showed a 15% higher accuracy than standalone LightGBM models in predicting customer churn, underscoring the advantage of incorporating sequential and attention-based processing for nuanced feature interactions.

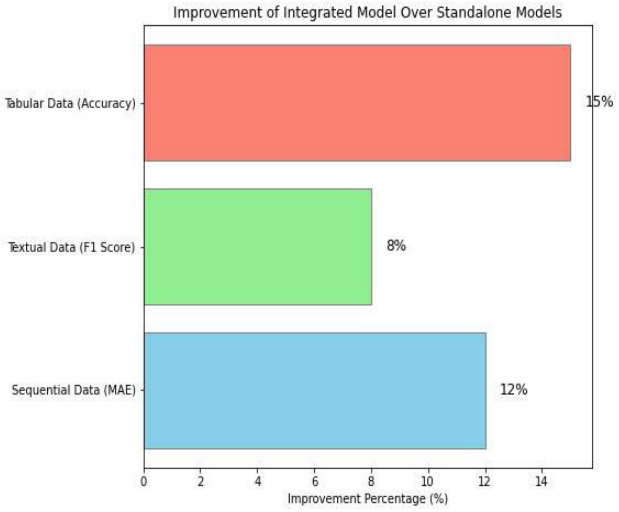


Figure 5. Performance of Evaluation matreces accuracy, F1 score, and MAE.

4.2 Comparative Analysis

The comparative analysis with baseline models indicates that while LSTM, Transformer, and LightGBM models excel in their respective domains; their integration offers a versatile and robust solution that harnesses the strengths of each. Notably, the integrated model's performance underscores the synergistic effect of combining sequential processing, attention mechanisms, and efficient gradient boosting.

Table 1: Performance Comparison of Predicative Models Across Datasets

Model	Dataset Type	Accuracy (%)	F1 Score	MAE
Integrated Model	Sequential	93.5	92.0	0.45
LSTM	Sequential	88.7	87.5	0.58
Transformer	Sequential	89.2	88.1	0.55
LightGBM	Sequential	85.3	84.7	0.62
Integrated Model	Texual	91.8	90.4	N/A
LSTM	Texual	87.0	86.2	N/A
Transformer	Texual	89.6	88.9	N/A
LightGBM	Texual	82.5	81.9	N/A
Integrated Model	Tabular	94.2	93.6	0.36
LSTM	Tabular	86.8	86.1	0.57
Transformer	Tabular	87.4	86.7	0.54
LightGBM	Tabular	90.3	89.7	0.41

Table 1 illustrates that the integrated model outperforms the standalone LSTM, Transformer, and LightGBM models in all evaluated metrics across sequential, textual, and tabular datasets.

4.3 Discussion

The results affirm the hypothesis that an integrated approach can significantly enhance predictive modeling capabilities across various data types. The integration not only addresses the limitations of each model when used in isolation but also introduces a flexible architecture that adapts to the nature of the dataset.

Sequential Data: The LSTM and Transformer synergy provides a more nuanced understanding of temporal dependencies, crucial for accurate forecasting.

Textual Data: The combination of LSTM's ability to process sequences and Transformer's attention mechanism enhances the model's capacity to understand and generate nuanced language interpretations.

Tabular Data: LightGBM's efficiency, when combined with the depth of understanding from LSTM and Transformer models, enables a more sophisticated analysis of structured data, leading to improved predictive performance.

These findings have significant implications for the development of predictive models capable of handling a wide range of data types with higher accuracy and efficiency. The integrated model not only broadens the applicability of machine learning solutions but also opens avenues for research into further optimization of hybrid architectures.

Implications for Future Research

The promising results of integrating LSTM, Transformer, and LightGBM models suggest several directions for future research:

Optimization of Model Integration: Exploring more sophisticated methods for integrating the models could further enhance performance. This includes the development of dynamic weighting mechanisms to adjust the contribution of each model based on the dataset.

Application to New Domains: Applying the integrated model to new domains, such as healthcare diagnostics or financial market prediction, could demonstrate its versatility and adaptability to different challenges.

Scalability and Efficiency: Future work could focus on improving the scalability and computational efficiency of the integrated model, making it more accessible for real-world applications with large datasets.

The integration of LSTM, Transformer, and LightGBM models represents a significant step forward in the field of predictive analytics. By leveraging the strengths of these diverse models, we can achieve a level of predictive accuracy and efficiency that surpasses what

any of the models could achieve independently. This research not only contributes to the theoretical understanding of model integration but also provides a practical framework that can be adapted and optimized for a wide range of applications.

V. CONCLUSION

This study introduced a novel integrated model combining Long Short-Term Memory (LSTM) networks, Transformer models, and Light Gradient Boosting Machine (LightGBM) to address the challenges of predictive modeling across various data types. The experimental results demonstrated that the integrated model significantly outperforms standalone implementations of LSTM, Transformer, and LightGBM models in tasks involving sequential, textual, and tabular data.The synergy achieved by combining these models highlights the potential for creating more versatile and powerful machine-learning solutions. Future research should focus on refining the integration mechanism, exploring applications in new domains, and enhancing model efficiency for larger datasets. This research presents a promising direction for advancing predictive analytics, underscoring the value of hybrid models in leveraging the strengths of diverse machine learning architectures. The integrated approach not only broadens the applicability of predictive models but also sets a foundation for future innovations in the field.

REFERENCES

[1] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735 1780.

[2] Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[3] Ke, G., et al. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

[4] Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. Neural computation, 12(10), 2451-2471.

[5] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. Advances in neural information processing systems, 27.

[6] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics.

[7] Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[8] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

[9] Graves, A. (2013). Generating sequences with recurrent neural networks. arXiv preprint arXiv:1308.0850.

[10] Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

[11] Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. IEEE transactions on neural networks and learning systems, 28(10), 2222-2232.

[12] Vaswani, A., et al. (2017). Attention is all you need. Advances in neural information processing systems, 30.

[13]Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics.

[14] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. International Conference on Learning Representations.

[15] Niu, F., Recht, B., Ré, C., & Wright, S. J. (2011). HOGWILD!: A lock-free approach to parallelizing stochastic gradient descent. Advances in neural information processing systems, 24.

[16] Guo, C., Berkhahn, F. (2016). Entity embeddings of categorical variables. arXiv preprint arXiv:1604.06737.

[17] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in Neural Information Processing Systems, 31.

[18] Siami-Namini, S., Tavakoli, N., & Namin, A. S. (2018). A comparison of ARIMA and LSTM in forecasting time series. 17th IEEE International Conference on Machine Learning and Applications (ICMLA).

[19] Khandelwal, U., He, H., Qi, P., & Jurafsky, D. (2020). Sharp nearby, fuzzy far away: How neural language models use context. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics.

[20] He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., ... & Candela, J. Q. (2014). Practical lessons from predicting clicks on ads at Facebook. Proceedings of the Eighth International Workshop on Data Mining for Online Advertising.

[21] Baldi, P., Sadowski, P., & Whiteson, D. (2014). Searching for exotic particles in high-energy physics with deep learning. Nature Communications, 5, 4308.

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

A PERFORMANCE STUDY OF TWO JPEG COMPRESSION APPROACHES

Doaa M. Elmourssi[1], Wahida A. Mansouri[1,2], Wiam A. Elyass[1], Salwa H. Othman[1,2], Somia Asklany[1,3]*

[1]Department of Computer Science and Information Technology, Faculty of sciences and arts, Turaif, Northern Border University, Arar 91431, Kingdom of Saudi Arabia

[2]LETI laboratory, University of Sfax, Tunisia

[3]Modern academy for science and technology, Maadi, Cairo, Egypt

**Corresponding author E-mail: doaa.elmourssi@nbu.edu.sa*

ABSTRACT

As technology continues to advance, and the transition into the digital age, we find ourselves dealing with an ever-expanding volume of information, often leading to challenges in management. Consequently, there is a growing need to store and retrieve digital information in a manner that is both efficient and effective. This paper discusses jpeg image compression techniques; and presents a comparison between two modifications of jpeg image compression techniques. Both of two modifications rely on applying DCT to the divided blocks of the image. They differ in the way of selecting the coefficients resulted from DCT to be quantized in order to remove redundancy. The quality of the compressed images is evaluated using Peak Signal to Noise Ratio (PSNR), Weighted Peak Signal to Noise Ratio (WPSNR) and spatial frequency measurement (SFM).

Keywords: jpeg compression, SFM, WPSNR, SSIM.

1. INTRODUCTION

With each passing day, computers are becoming increasingly powerful, leading to a surge in the utilization of digital images. However, this widespread use of digital images brings about a significant challenge managing the substantial volume of data these images represent. The storage and transmission of uncompressed multimedia data, encompassing graphics, audio, and video, pose considerable demands on storage capacity and bandwidth, highlighting a pressing issue in the digital landscape. Reducing the bandwidth needs of any given device will result in significant cost reductions and will make use of the device more affordable.

Image compression plays a significant role in the realm of image information processing. The dimensions of image data hold paramount importance during image processing as they can

impact various aspects, including design architecture, memory requirements, processing time for large datasets, and overall processing complexity. In the work [10] K. Dharavath and S. Bhukya introduce a novel method for compressing images while maintaining their essential features intact. Recent research such as work in [11] aims to Identify Image Modifications using DCT and JPEG Quantization Technique. in [12] the authors propose a new approach which uses the features of three methods: Discrete HAAR Wavelet Transform (DHWT), the Singular Value Decomposition (SVD) and Joint Photographic Experts group (JPEG). In [14]-[15] the authors present A Review of different image Compression Techniques. In the work [16] Hussain, A, Al-Fayadh, A and Radi, N introduce Image Compression Techniques: A Survey in Lossless and Lossy algorithms. Indeed, they begin by compressing the image by deleting some value from the input image.

Data compression offers ways to represent data in a more compact way, so that one can store more data and transmit it faster. The advantages of data compression come at the expense of numerical computations, and therefore we can trade off computations for storage or bandwidth. Before storing or transmitting data we process it in such a way that will require fewer bits for its representation [2]. One of the most popular and comprehensive still farm compression is the jpeg (for joint photographic expert group) standard in the baseline coding system, which is based on discrete cosine transform [3] and is adequate for most compression applications.

This paper is organized as follows: Section II explains the principles of compression. Section III presents the proposed algorithm. IV briefly explains the performance evaluation criteria. Section V introduces the experimental results and section VI gives the concluding remarks.

2. PRINCIPLES BEHIND COMPRESSION

Recently, most of the proposed localization schemes are based on RSSI technique but the RSSI signal propagation models easily suffer from outer uncertain influences, such as signal fading, non-uniform spreading, and reflections. An RSSI-based approach therefore needs more data than other methods to achieve higher accuracy. Data compression techniques exploit inherent redundancy and irrelevancy by transforming a data file into a smaller file from which the original image file can later be reconstructed, exactly or approximately [1].

Redundancy reduction focuses on eliminating duplications present in the signal source, such as image or video data. Irrelevancy reduction, on the other hand, involves discarding portions of the signal that are unlikely to be perceptible to the signal receiver, specifically the Human Visual System (HVS). In general, there are four types of redundancy can be summarized:

- Spatial Redundancy due to correlation between neighboring pixels.
- Spectral redundancy arises from the correlation between distinct color planes or spectral bands.
- Temporal redundancy due to correlation between adjacent frames in a sequence of images (in video applications).
- Statistical Redundancy due to statistical properties of images.

Image compression research aims at reducing the number of bits needed to represent an image by removing the spatial and statistical redundancies as much as possible [15]. The volume of data required describing such images greatly slow transmission and makes storage prohibitively costly. The information contained in images must, therefore, be compressed by extracting only visible elements, which are then encoded [14].

The volume of data is significantly diminished through this process. The primary objective of image compression is to decrease the bit rate for transmission or storage while preserving an acceptable level of fidelity or image quality. Compression can be broadly categorized in two

ways:

- Lossless Compression: This method compresses data in a way that allows for complete reconstruction (uncompressing) without any loss of detail or information [16].
- Lossy Compression: In contrast, lossy compression does not maintain the exact pixel-to-pixel representation of the original image. Instead, it capitalizes on the limitations of the human eye to approximate the image in a way that appears visually identical to the original. While lossy methods can achieve significantly higher compression rates than lossless methods, they need to be employed judiciously to avoid noticeable degradation in image quality Processing Steps for jpeg Image Compression [17].

3. PROCESSING STEPS FOR JPEG IMAGE COMPRESSION

3.1 JPEG Encoder

The encoding process is started by dividing image data into square blocks and applying DCT show in figure 1. The resultant coefficients are selected and quantized using two approaches [3]-[4].

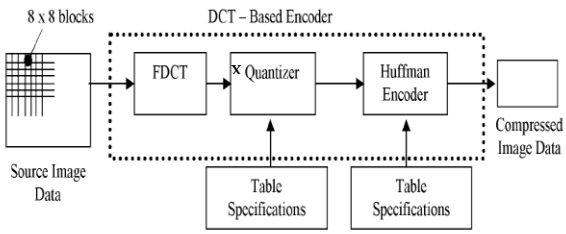


Figure:1 DCT-Based JPEG Encoder Processing Steps

Step1: The image is divided into non overlapping 8*x*8 blocks of pixels start from left to right, top to bottom.

Step2: Every pixel in the block is shifted from an unsigned integer with range $[0, 2^p-1]$ to a signed integer with range $[-(2^{p-1}), 2^{p-1}-1]$ by subtracting 2^{p-1} from the value where p is the number of bits per channel (or bits per value). In the case of the standard 8-bit channel the numbers are shifted from [0, 255] to [−128.127] by subtracting 128.Beacause DCT is designed to work in this range.

Step3: Each shifted pixel in the 8x8 block is then transformed into frequency domain via discrete cosine transform (DCT). The transformed 8x8 block now consists of 64 DCT coefficients. The first coefficient (0, 0) is the DC component in the block and the other 63 coefficients are AC component of the block

Step4: The resultant transformed coefficients are then passed to the quantizer which simply reduces the number of bits needed to store the transformed coefficients by reducing the precision of these values. Instead of passing all coefficients to the quantizer, only some selected coefficients (x) will be utilized. Two approaches for selection will be adopted.

- 1st approach: after sorting coefficients in all blocks, only the top k largest coefficients in magnitude will be kept and the other coefficients will be set to zeros.
- 2nd approach: instead of applying the sorting globally, coefficients of in each block will be sorted locally in decreasing order and only the top *(k/c)* largest coefficients in magnitude will be kept where c denotes the number of blocks. This process will be repeated for all blocks.

Step 5: The selected coefficients will be passed through the quantizer. Each of the 64 DCT coefficient F(u,v) is quantized using the stander quantization matrix Q(u,v) by dividing the each coefficient by the corresponding quantize element in the quantization matrix and rounded to the nearest integer as

$$F_q(u,v) = Round\left(\frac{F(u,v)}{Q(u,v)}\right) \qquad (1)$$

Quality of the reconstructed image can be controlled by a user by selecting a quality level. The value of quality level may vary from 1 to 100 if another level of quality is desired. It is permissible to use scalar multiples of the JPEG standard quantization matrix. For quality<50 (more compression, lower image quality), the standard quantization matrix is multiplied by 50/quality level and for quality level>50 (less compression, more image quality), the standard quantization matrix is multiplied by (100-quality level)/ 50.

Step 6: After quantization of the DCT coefficients the quantized DC-coefficients are

treated differently than the quantized AC-coefficients. The processing order of all coefficients is specified by the zig-zag sequence shown in Figure 2. The differences between successive DC-coefficients are very small values. Thus, each DC-coefficient is encoded by subtracting the DC-coefficient of the previous data unit, as shown in Figure 3, and subsequently using only the difference.

The DCT processing order of the AC-coefficients using the zig-zag sequence illustrates that coefficients with lower frequencies (typically with higher values) are encoded first, followed by the higher frequencies (typically zero or almost zero) show in Figure 2. The result is an extended sequence of similar data bytes, permitting efficient entropy encoding.

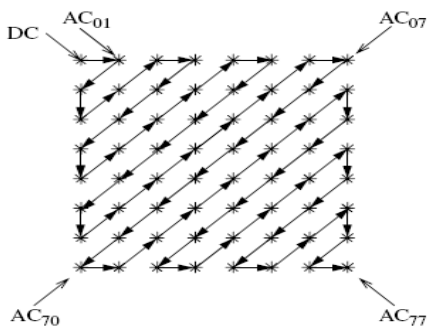


Figure 2: Zig-zag ordering of AC coefficients

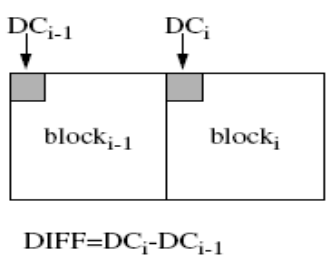


Figure 3: differential coding of DC

Step7: Finally, the last block in the JPEG encoder is the entropy coding, which provides additional compression by encoding the quantized DCT coefficients into more compact form show in figure 3. The JPEG standard specifies two entropy coding methods: Huffman coding and arithmetic coding [5]. The baseline sequential JPEG encoder employs Huffman coding. The Huffman coder converts the DCT coefficients after quantization into a compact binary sequence using two steps: (1) forming intermediate symbol sequence, and (2) converting intermediate symbol sequence into binary sequence using Huffman tables.

3.2 JPEG Decoder

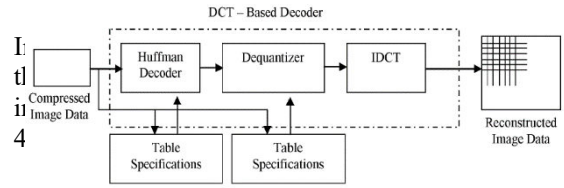


Figure 4: DCT-Based JPEG Decoder Processing Steps

Step1: First, an entropy decoder (such as Huffman) is applied to the compressed image data. The binary sequence is converted to symbol sequence using Huffman tables, and then the symbols are converted into DCT coefficients.

Step2: The dequantization is applied to the resultant coefficients using the following equation:

$$F_q(u,v) = F_q(u,v).Q(u,v) \qquad (2)$$

Step3**:** Then, the Inverse Discrete Cosine Transform (IDCT) is applied to the dequantized coefficients in order to convert the image from frequency domain into spatial domain.

4. IMAGE QUALITY MEASUREMENTS

Assessing image quality is crucial in different image processing applications. After creating and applying an image compression system, it becomes essential to assess its performance. This evaluation should enable a comparison of results with other image compression techniques. Image quality metrics offer a way to gauge the similarity between two digital images by leveraging variations in the statistical distribution of pixel values [9].

4.1 Peak Signal to Noise Ratio (PSNR)

It serves as a commonly employed measure of precision. A low Peak Signal to Noise Ratio (PSNR) indicates low image quality. The definition of PSNR is as follows:

$$PSNR = 10\log\left(\frac{L^2}{MSE}\right) \quad (3)$$

Where $L = 255$ is the dynamic range of the pixel values.
And

$$MSE = \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[y(i,j) - x(i,j)]^2 \quad (4)$$

4.2 Weighted Peak Signal to Noise Ratio (WPSNR)

The Weighted Peak Signal to Noise Ratio (WPSNR) is an extension of the conventional PSNR, where each PSNR term is weighted by a local "activity" factor associated with the local variance [7].

$$WPSNR = 10\log\left(\frac{L^2}{\|(y-x).NVF\|^2}\right) \quad (5)$$

where

$$NVF = 1/(1+\theta\sigma_x^2(i,j)) \quad (6)$$

$$\sigma_x^2(i,j) = \frac{1}{(2L+1)^2}\sum_{m=-L}^{L}\sum_{n=-L}^{L}(x(i+m,j+n) - \bar{x}(i,j))^2 \quad (7)$$

$$\theta = \frac{D}{\sigma_{x\,max}^2} \quad (8)$$

where σ^2_{xmax} is the maximum local variance of a given image and $D \in [50,150]$ is a determined parameter.

4.3 Maximum Difference (MD)

A high Maximum Difference (MD) value indicates low image quality. MD is defined as follows:

$$MD = \max(|y(i,j) - x(i,j)|) \quad (9)$$

4.4 Correlation Coefficient

The correlation coefficient serves as a prevalent metric to quantify the similarity between two images. The correlation coefficient can take values between -1 and 1, with a stronger correlation approaching values closer to -1 or 1. The computation is carried out using the subsequent equation:

$$CF = \frac{\sum_{i,j}[y(i,j) - \bar{y}][x(i,j) - \bar{x}]}{\sqrt{\sum_{i,j}[y(i,j) - \bar{y}]^2 \sum_{i,j}[x(i,j) - \bar{x}]^2}} \quad (10)$$

4.5 Structural Similarity Based Metrics

Another category of image quality measure is based on the assumption that the human visual system is highly adapted to extract structural information from the viewing field [8]. The error sensitivity approach estimates perceived errors to quantify image degradations, while this approach considers image degradations as perceived structural information variation. The structural Similarity *(SSIM)* index can be calculated as a function of three components: luminance, contrast and structure.

$$SSIM(x,y) = [l(x,y)]^{\alpha} \cdot [c(x,y)]^{\beta} \cdot [s(x,y)]^{\gamma} \quad (11)$$

This results in a specific form of the SSIM index:

$$SSIM(x,y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)} \quad (12)$$

where
$C_1 = (K_1 L)^2$, $K_1 << 1$ and $C_2 = (K_2 L)^2$, $K_2 << 1$.

5. EXPERIMENTAL RESULTS

Six test images (512×512, 8 bits/pixel) with different spatial and frequency characteristics, as shown in Figure 5: Lena, Baboon, peppers, Goldhill, Boat and Barb are used. Characteristics of test images are evaluated in spatial domain using spatial frequency measure (SFM) [6].

Figure 5. Image Database (size of 512×512

The spatial frequency measurement (SFM) indicates the overall activity level in an image. SFM is defined as follow:

$$SFM = \sqrt{(R)^2 + (C)^2} \quad (13)$$

$$R = \sqrt{\frac{1}{MN}\sum_{m=1n=2}^{M,N}[x(m,n) - x(m,n-1)]^2} \quad (14)$$

$$C = \sqrt{\frac{1}{MN}\sum_{n=1,m=2}^{N,M}[x(m,n) - x(m-1,n)]^2} \quad (15)$$

Where R is row frequency, C is column frequency, x (m, n) denotes the samples of image, M and N are number of pixels in row and column directions, respectively. The large value of *SFM* means that image contain component in high frequency area [13].

a. Measuring Spatial Frequency (SFM)

The Spatial frequencies (SFM) and values computed for the above set of images are given in Table 1.

Test image Baboon has a lot of details and consequently large SFM. Large value of SFM means that image contains components in high frequency area. It returns that Baboon presents.

TABLE 1. spatial frequency measure of images

	Lena	Baboon	Peppers	Goldhill	Boat	Barb
SFM	14.042	36.5146	15.8446	16.1666	17.8565	29.4567

b. Evaluating Perceptual Quality

PSNR, wPSNR, MD, CF and SSIM values of each of the two approaches of JPEG compression are recorded in Table 2, 3,4,5, and 6 respectively. All values are recorded at *k*=8192

TABLE 2. PSNR values (in dB) of the images.

Method	Lena	Baboon	Peppers	Goldhill	Boat	Barb
Apr-1	32.2410	30.3528	31.5590	31.2419	31.8917	31.2373
Apr-2	34.5701	30.4406	34.1011	32.2553	33.6249	31.9630

In Table 2 we record the PSNR values of each of the two approaches of JPEG compression at k=8192. We note that the second approach provides better PSNR than the first one at the same *k*.

TABLE 3 wPSNR values (in dB) of the images

Method	Lena	Baboon	Peppers	Goldhill	Boat	Barb
Apr-1	133	147	185	113	153	142
Apr-2	92	147	130	88	109	108

TABLE 4 MD values of the images

Method	Lena	Baboon	Peppers	Goldhill	Boat	Barb
Apr-1	27.249	21.2859	26.4219	27.0946	25.8598	24.0606
Apr-2	31.0313	22.1589	30.5118	28.7770	28.5342	25.9047

The wPSNR values for each of the two JPEG approaches at k=8192 are listed in Table 3. We observe that at the same k, the second approach offers better wPSNR than the first.

In Table 4 we record the MD values of each of the two approaches of JPEG compression at k. we note that the second approach provides better MD than the first one at the same *k*.

TABLE 5 Correlation coefficient values of the images

Method	Lena	Baboon	Peppers	Goldhill	Boat	Barb
Apr-1	0.9728	0.8542	0.9773	0.9735	0.9687	0.9562
Apr-2	0.9887	0.8826	0.9912	0.9821	0.9832	0.9716

The CR values for each of the two JPEG approaches at k=8192 are listed in Table 5. We observe that at the same k, the second approach offers better wPSNR than the first.

TABLE 6 Structural Similarity values of the images

Method	Lena	Baboon	Peppers	Goldhill	Boat	Barb
Apr-1	0.7852	0.5265	0.7634	0.6922	0.7572	0.7171
Apr-2	0.8223	0.5279	0.7950	0.7128	0.7928	0.7309

Table 6 lists the SSIM values for the two JPEG approaches at k=8192. We find that the second

approaches provide better SSIM than the first at the same k.

All above tables assure that the 2nd approach provides better image quality than the 1st one at the same *k*. Moreover, the 2nd approach is quite faster compared to that of 1st one. Figure 6 and figure 7 show the original and compressed images using different approaches at different values of *k*.

Figure 6. (a)Original image, (b),(c) compressed images using 1st and 2nd approach respectively at *k*=8192

Figure 7. (a)Original image, (b),(c) compressed images using 1st and 2nd approach respectively at *k*=163840

A comparison between the two approaches is presented by measuring the PSNR, wPSNR, MD, CF and SSIM for the compressed images using the two approaches at different values of *k*.

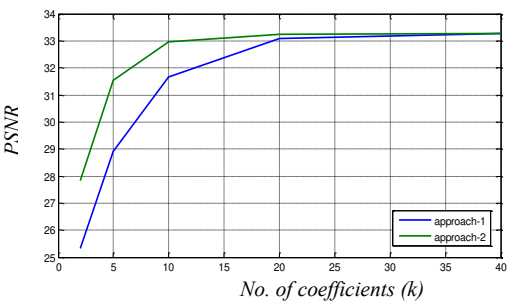


Figure 8: PSNR versus different values of k

Figure 8 represents the relation between the PSNR and number of selected coefficients for each of the two JPEG approaches. We note that the second approaches provide better PSNR than the first at the same k.

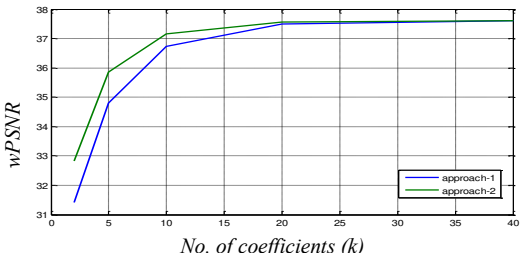


Figure 9: wPSNR versus different values of k

Figure 9 illustrates the relation between the wPSNR and number of selected coefficients for each of the two JPEG approaches. We note that the second approaches provide better wPSNR than the first at the same k.

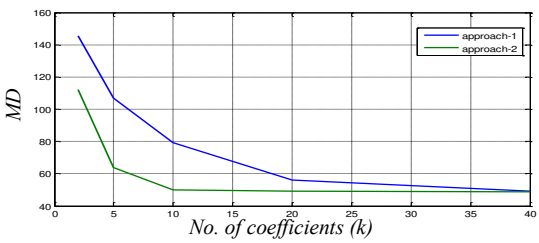


Figure10. MD versus different values of *k*

Figure 10 shows the relation between the MD and number of selected coefficients for each of the two JPEG approaches. We find that the second approaches provide better wPSNR than the first at the same k.

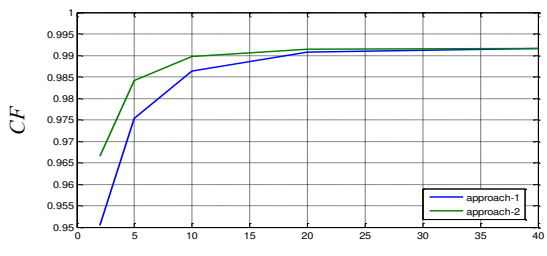


Figure 11.CF different values of k

Figure 11 represents the relation between the CF and number of selected coefficients for each of the two JPEG approaches. We note that the second

approaches provide better wPSNR than the first at the same k.

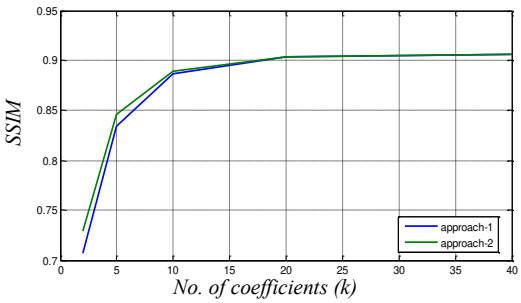


Figure 12. SSIM versus different values of k

Figure 12 illustrate the relation between the SSIM and number of selected coefficients for each of the two JPEG approaches. We remark that the second approach provides better wPSNR than the first at the same k.

From all the previous figures we assure that the second approaches provide better image quality than the first approaches in the same number of selected coefficients

5. CONCLUSION

This paper clarifies the advantage of choosing the coefficients to be quantized globally in achieving better perceptual quality. The experimental results emphasis the better quality of the compressed images resulted from the approach relying on quantizing the largest coefficients selected globally than the images resulted from the approach relying on quantizing the largest coefficients selected locally in each block. While JPEG is primarily known for lossy compression, future advancements might focus on improving lossless compression capabilities. This would be particularly beneficial for applications requiring perfect image reconstruction without any loss of information. As data security and privacy become increasingly important, future JPEG compression techniques may incorporate encryption and watermarking capabilities to protect sensitive information within images.

REFERENCES

[1] Ze-NianLi,Mark S.Drew "Fundamentals of Multimedia" Prentice Hall, INC., 2003.

[2] Pennebaker, William B. and Joan L. Mitchell, JPEG: Still Image Data Compression Standard, Van Nostrand Reinhold, New York, 1993.

[3] O. Rippeland , Bourdev,“Real-time adaptive image compression”, International Conference on Machine Learning, Sydney, Australia, 2017

[4] Wallace, Gregory K. The JPEG Still Picture Compression Standard paper Submitted in December 1991 for publication in IEEE Transactions on Consumer Electronics.

[5] D.A. Huffman," A Method for the Construction of Minimum-Redundancy Codes" Proc.IRE 40,1098-1101(1952).

[6] M. Eskicioglu, P. S. Fisher, Image Quality Measures and Their Performance, IEEE Transactions on Communications, Vol. 43, No. 12, December 1995, pp. 2959-2965

[7] S. Voloshynovskiy, S. Pereira, A. Herrigel, N. Baumgartner, T. Pun, “A Stochastic Approach to Content Adaptive Digital Image Watermarking”, Proceedings of the Third International Workshop on Information Hiding, pp.211-236, 1999.

[8] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error measurement to structural similarity" IEEE Transactions on Image Processing, vol. 13, No. 1, January 2004.

[9] Rafael C. Gonzalez, Richard E. Woods; "Digital Image Processing", Edition 4,2018, page 1022

[10] K. Dharavath and S. Bhukya, "A Novel Approach for Improving Image Compression Ratio," *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India,

2022, pp. 1-4, doi: 10.1109/GCAT55367.2022.9972021.

[11] P. Kubal, N. Pulgam and V. Mane, "Identifying Image Modifications using DCT and JPEG Quantization Technique," *2023 IEEE 8th International Conference for Convergence in Technology (I2CT)*, Lonavla, India, 2023, pp. 1-5, doi: 10.1109/I2CT57861.2023.10126200.

[12] R. Ranjan, P. Kumar, K. Naik and V. K. Singh, "The HAAR-the JPEG based image compression technique using singular values decomposition," 2022 2nd International Conference on Emerging Frontiers in Electrical and Electronic Technologies (ICEFEET), Patna, India, 2022, pp. 1-6, doi:10.1109/ICEFEET51821.2022.9848400.

[13] Doaa Mohammed, Fatma Abou-Chadi (2011), Image Compression Using Block Truncation Coding, Cyber Journals: Multidisciplinary Journals in Science and Technology, Journal of Selected Areas in Telecommunications (JSAT), February Edition.

[14] Garima Garga, Raman Kumarb (2022), Analysis of Different Image Compression Techniques: A Review, International Conference on Innovative Computing & Communication (ICICC) 2022.

[15] Surabhi N, 2 Sreeleja N Unnithan (2017), Image Compression Techniques: A Review, International Journal of Engineering Development and Research (IJEDR) ISSN: 2321-9939

[16] Hussain, A, Al-Fayadh, A and Radi, N (2018) Image Compression Techniques: A Survey in Lossless and Lossy algorithms. Neurocomputing. ISSN 0925-2312

[17] Sarkar, J. B., Poolakkachalil, T. K., & Chandran, S. (2018). Novel Hybrid Lossy Image Compression Model using Run-Length Coding and Huffman Coding. International Journal of Computer Science and Information Security (IJCSIS), 16(10), 103-107.

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

DECODING THE DECEPTION: A COMPREHENSIVE ANALYSIS OF CYBER SCAM VULNERABILITY FACTORS

Asma A. Alhashmi[1], Huda K. Sheatah[1], Imen B. Mohamed[1], Chams Jabnoun[1], Firas M. Allan[1], Aida Dhibi[1], Doaa M. Elmourssi[2], Abdulbasit A. Darem[1]*

[1]*Department of Computer Science, Faculty of Science, Northern Border University*
[2]*Faculty of sciences and arts, Turaif, Northern Border University, Arar 91431, Kingdom of Saudi Arabia.*
huda.sheatah@nbu.edu.sa, eman.bnmohmmed@nbu.edu.sa, chams.sallami@nbu.edu.sa, firas.allan@nbu.edu.sa, aaedah.alsagheer@nbu.edu.sa, doaa.elmourssi@nbu.edu.sa

ABSTRACT:

This paper presents a comprehensive analysis of the multifaceted factors influencing susceptibility to cyber scams. This study delves into the complexities surrounding individuals' susceptibility to cyber scams, integrating psychological, behavioral, technical, and environmental perspectives to offer a nuanced understanding of digital deception vulnerabilities. It highlights the exploitation of cognitive biases and emotional vulnerabilities by scammers, discusses the impact of habitual online behaviors and security fatigue, and addresses the challenges posed by the rapid evolution of cyber threats. Additionally, it explores societal and cultural influences on scam susceptibility. Proposing an integrated prevention framework, this research emphasizes a multifaceted approach encompassing education, technological solutions, policy development, and psychological interventions to mitigate the risks and impacts of cyber scams. Our investigation into cyber scam susceptibility unravels the intricate interplay of psychological, behavioral, technical, and environmental factors shaping individuals' vulnerabilities. It reveals that susceptibility is not solely the product of individual ignorance or oversight but results from a complex mesh of human psychology, habitual behaviors, technological advancements, and socio-cultural influences. The proposed comprehensive framework for combating cyber scams underscores the necessity for a collaborative, interdisciplinary approach that combines educational initiatives, policy reforms, technological advancements, and psychological support. Future research should aim at refining this framework, focusing on the dynamic and evolving nature of cyber scams to devise effective, adaptive strategies for prevention and intervention, ensuring a safer digital landscape for users worldwide.

Keywords: Cyber Scams, Psychological Vulnerability, Behavioral Security, Digital Deception, Cybersecurity Awareness.

I. INTRODUCTION

The proliferation of digital technologies has revolutionized various aspects of modern life, offering unprecedented opportunities for businesses, communication, and leisure activities (Eze et al., 2023). However, this digital transformation has also given rise to a myriad of cyber threats, including scams and fraudulent activities, which pose significant challenges to individuals, organizations, and societies at large. Recent research has provided valuable insights into the multifaceted factors influencing susceptibility to cyber scams, encompassing technical, non-technical, and organizational dimensions. This introduction aims to provide a comprehensive overview of the factors influencing susceptibility to cyber scams, drawing on recent studies to elucidate the diverse elements shaping vulnerability to online fraudulent activities. Technical Vulnerability Factors The effects of cyber-attacks are particularly critical for

technology startups, as they often possess low cybersecurity maturity levels, making them more susceptible to malicious activities in the digital realm (Marican et al., 2023). Furthermore, the increasing connectivity of digital and cyber-physical systems has necessitated heightened attention to cybersecurity to enhance the integrity, confidentiality, and availability of data, underscoring the technical vulnerability factors associated with the evolving cyber ecosystem (Angelelli et al., 2023). Additionally, the development of advanced cyber security systems based on anomaly detection using Artificial Neural Networks has become imperative in addressing the escalating internet crimes and enhancing cybersecurity (Hephzipah et al., 2023). Non-Technical Influences on Susceptibility Beyond technical aspects, individual differences in susceptibility to cyber scams have been a focal point of recent research. Studies have delved into victims' shock absorption mechanisms in response to cybercrime, shedding light on the psychological and emotional dimensions of susceptibility to online scams (Eze et al., 2023). Moreover, the design of an effective organizational culture has been identified as a crucial non-technical measure to guard against the cyber risks posed by emerging technologies, emphasizing the significance of non-technical influences on vulnerability factors (Watkins, 2023). Furthermore, the theoretical basis and occurrence of internet fraud victimization have been explored, highlighting the role of objective knowledge and experience in specific fields in shaping susceptibility to online fraudulent activities (Shang et al., 2023). Organizational Resilience and Cybersecurity Strategies Organizational cyber resilience has emerged as a critical factor in mitigating vulnerability to cyber scams. Research has emphasized the need for an effective organizational culture to guard against cyber risks, particularly in the context of emerging technologies, underscoring the organizational dimension of susceptibility to cyber scams (Watkins, 2023). Additionally, the quantitative assessment of the relative impacts of different factors on susceptibility modeling has provided valuable insights into the organizational and environmental influences on vulnerability to cyber threats, offering a holistic perspective on susceptibility factors (Khaldi et al., 2023).

In the digital era, the proliferation of cyber scams represents a significant threat to individuals and organizations worldwide. These scams, characterized by their deceptive and manipulative tactics, exploit vulnerabilities across various dimensions - psychological, behavioral, technical, and environmental. The sophistication of these scams has grown, paralleling advancements in technology and the increasing reliance of individuals on digital platforms. This introduction delves into the multifaceted nature of cyber scams, exploring how they leverage human psychology, behavioral patterns, technical gaps, and environmental factors to ensnare victims. The psychological dimension is crucial in understanding why individuals fall prey to cyber scams. Scammers expertly manipulate cognitive biases and emotional responses. Trust, greed, fear, and the desire for social connection are key psychological factors that scammers exploit. Drawing on theories from psychology and behavioral economics, we examine how cognitive heuristics and emotional responses can lead to poor decision-making, making individuals susceptible to scams. The role of social engineering in phishing attacks, which plays on trust and authority, is a prime example of this exploitation. The psychological impact is profound, often leaving victims with long-lasting emotional and financial scars. Behavioral factors play a significant role in susceptibility to cyber scams. Habitual behaviors, such as routine responses to emails or social media interactions, can become vulnerabilities. The concept of 'security fatigue' is critical here; repeated exposure to security warnings and protocols can lead to complacency. This section discusses how habitual online behaviors, when unexamined, can make individuals more vulnerable to sophisticated phishing attacks and social engineering tactics. The importance of cultivating cyber hygiene practices, such as regularly updating passwords and scrutinizing email sources, is emphasized as a countermeasure to these behavioral vulnerabilities. Technical knowledge and skills are a double-edged sword in the realm of cyber scams. A lack of technical understanding can leave individuals exposed to complex scams, while overconfidence in one's technical abilities can lead to underestimating the sophistication of scammers. This section explores the technical complexities of modern cyber scams, including malware, ransomware, and advanced phishing techniques. It also discusses the importance of cybersecurity education and awareness as critical tools in combating these threats. The role of technological solutions, such as antivirus software and firewalls, is acknowledged, but the need for continuous education and vigilance is underscored, given the ever-evolving nature of cyber threats. The environmental dimension encompasses the broader social and institutional contexts that shape an individual's susceptibility to cyber scams. Cultural norms, social influences, and institutional policies can either mitigate or exacerbate the risk of falling victim to scams. This section examines the role of social networks in spreading scams and how institutional policies and regulations can help create a safer cyber environment. The influence of peer groups and social

media in shaping online behaviors and perceptions towards scams is discussed, highlighting the need for community-based awareness and education programs. In conclusion, the interplay between these dimensions collectively contributes to the risk of falling victim to cyber scams. A holistic approach is essential in addressing these threats, combining psychological insights, behavioral interventions, technical solutions, and environmental strategies. The future of cybersecurity lies in understanding and addressing these multifaceted vulnerabilities, fostering a culture of awareness and resilience against the ever-present threat of cyber scams.

II. RELATED WORK

The psychological dimension plays a crucial role in understanding susceptibility to cyber scams. Research in this area focuses on cognitive biases, emotional manipulation, and the psychological profiles of victims. Whitty and Buchanan (2012) delve into the psychological manipulation used in online romance scams, highlighting how scammers exploit victims' emotional vulnerabilities. Buchanan and Whitty (2014) further explore this by examining the psychological characteristics that make individuals more susceptible to these scams, such as loneliness and risk-taking behavior. Behavioral patterns significantly influence individuals' responses to cyber threats. Workman (2008) discusses how habitual behaviors and 'security fatigue' can lead to increased vulnerability to phishing attacks and social engineering. Crossler et al. (2013) extend this discussion by examining how individuals' routine activities and online behaviors, such as frequent online shopping or social media use, can increase their exposure to cyber scams. Technical knowledge and skills are critical in understanding and preventing cyber scams. Parsons et al. (2014) emphasize the importance of cybersecurity education in enhancing individuals' ability to recognize and respond to cyber threats. They argue that a lack of technical understanding can leave individuals vulnerable to more sophisticated scams, such as those involving malware or ransomware. The environmental dimension, including social and institutional factors, shapes individuals' susceptibility to cyber scams. Button et al. (2014) explore how social networks and cultural norms can influence individuals' perceptions and responses to cyber scams. They highlight the role of institutional policies and regulations in creating safer cyber environments and reducing the prevalence of scams. An interdisciplinary approach is essential in understanding and combating cyber scams. Research in this area combines insights from psychology, behavioral science, information technology, and social sciences. Moore et al. (2019) provide a comprehensive overview of this interdisciplinary approach, discussing how different fields contribute to a more holistic understanding of cyber scams and their prevention.

Hephzipah et al. (2023) developed a system for anomaly detection in cybersecurity using Artificial Neural Networks. Although the population and specific methodology were not detailed, the study represents a significant step forward in system development for real-time threat detection. Marican et al. (2023) systematically reviewed cybersecurity maturity frameworks for startups, emphasizing the unique needs of emerging technology companies. Their literature review suggests that startups must adopt tailored cybersecurity strategies to safeguard their operations and data. Shang et al. (2023) explored the theoretical basis of internet fraud victimization, examining decision-making processes that lead to victimhood. Their analysis contributes to a better understanding of how individuals become targets of internet fraud. Srivastava et al. (2023) investigated the impact of perceived value on online purchase intentions through empirical research. This study sheds light on consumer behavior in the digital marketplace, offering insights into how online retailers can enhance consumer trust and security perceptions. Angelelli et al. (2023) developed a theoretical framework for cyber-risk prioritization based on risk perception and decision-making. Using regression models, the study provides a novel approach to managing cyber risks in an increasingly complex digital environment. Ashwini et al. (2023) implemented an intrusion detection model using Support Vector Machine (SVM) techniques. This model addresses the challenge of identifying various types of cyberattacks, contributing to the development of more resilient cybersecurity systems.

Kim & Song (2023) measured cyber risk in financial and non-financial sectors using LDA and GARCH models. Their statistical analysis offers a quantitative approach to assessing cyber risk, providing valuable insights for risk management strategies. Tudosi et al. (2023) highlighted security weaknesses in distributed firewalls through penetration testing. Their security audit underscores the need for continuous vulnerability assessments to protect network infrastructures from emerging threats. Darem et al. (2023) classified cyber threats and countermeasures in the banking and financial sector, offering a comprehensive review of challenges and solutions in protecting financial data and operations from cybercriminals.

These studies collectively underscore the multifaceted nature of cybersecurity, highlighting the need for

continuous innovation, interdisciplinary approaches, and collaboration among stakeholders to address the complex challenges posed by cyber threats. As we advance, integrating insights from diverse research areas will be crucial in developing more effective cyber defense mechanisms and fostering a secure digital ecosystem. Looking forward, research in the field of cyber scams is moving towards more integrated and holistic approaches. Future studies are likely to focus on developing comprehensive models that incorporate psychological, behavioral, technical, and environmental factors. The aim is to develop more effective prevention and intervention strategies that address the multifaceted nature of cyber scams.

Table 1 provides a comprehensive overview of the studies, summarizing their focus areas, methodologies, populations, and key findings. In the next section, we will synthesize the findings from these studies, identifying common themes, methodological approaches, and gaps in the research. This analysis aims to highlight the evolution of cyber threat understanding and the varied approaches used to address these issues, from the psychological aspects of online scams to the technical solutions for cybersecurity threats. The increasing prevalence of online activities has led to a rise in cybersecurity threats, including online scams, fraud, and cyber-attacks on organizational infrastructure. This analysis reviews seminal works from literature, spanning topics from online romance scams to cybersecurity system development. By examining these studies, we aim to understand the multifaceted nature of cyber threats and the strategies developed to mitigate them.

Table 1. Comprehensive overview of the related studies

Study Reference	Focus Area	Variables Used	Tools	Methodology	Population	Key Findings
Whitty & Buchanan, 2012)	Online Romance Scam	Psychological traits, victim responses	Surveys, Interviews	Qualitative Analysis	Online Dating Users	Identified emotional vulnerabilities exploited in romance scams.
Buchanan & Whitty, 2014	Online Dating Scam	Victimhood causes, consequences	Surveys, Psychological Analysis	Quantitative & Qualitative Analysis	Online Dating Scam Victims	Explored psychological impact and characteristics of scam victims.
Workman, 2008)	Phishing and Social Engineering	Behavioral patterns, security awareness	Theoretical Framework	Theory-Grounded Investigation	General Internet Users	Highlighted the role of habitual behaviors in susceptibility to phishing.
Crossler et al 2013)	Behavioral Information Security	Online behaviors, security practices	Surveys, Statistical Analysis	Empirical Study	Internet Users	Discussed future directions for behavioral information security research.
Parsons et al. 2014)	Cybersecurity Awareness	Employee awareness, cybersecurity knowledge	HAIS-Q Questionnaire	Survey Analysis	Employees in Various Sectors	Determined the level of employee awareness in cybersecurity.
Button et al., 2014)	Impact of Fraud	Fraud impact on victims	Interviews, Case Studies	Qualitative Analysis	Fraud Victims	Analyzed the personal and familial impact of fraud.
Moore et al., 2019)	Economics of Online Crime	Online crime economics	Economic Analysis	Theoretical Study	Not Applicable	Discussed the economic perspective of online crime.
Hephzipah et al., 2023)	Cybersecurity System	Anomaly detection, network traffic	Artificial Neural Network	System Development	Not Specified	Developed a system for anomaly detection in cybersecurity.
Marican et al 2023)	Cybersecurity Maturity in Startups	Maturity assessment, startup needs	Literature Review	Systematic Review	Technology Startups	Reviewed cybersecurity maturity frameworks for startups.
Shang et al., 2023)	Internet Fraud Victimisation	Decision-making, victimization	Theoretical Framework	Theoretical Analysis	Internet Fraud Victims	Discussed the theoretical basis of internet fraud victimization.
Srivastava et al., 2023)	Online Purchase Intention	Perceived value, consumer behavior	Consumer Study	Empirical Research	Online Consumers	Studied the impact of perceived value on online purchases.
Angelelli et al., 2023)	Cyber-risk Perception	Risk perception, decision-making	Regression Models	Theoretical Study	Not Specified	Developed a framework for cyber-risk prioritization.
Ashwini et al 2023)	Intrusion Detectio Model	Cyberattack types, network traffic	Support Vector Machine	Model Implementation	Not Specified	Implemented an intrusion detection model using SVM
Kim & Song, 2023)	Cyber Risk Measurement	Loss data, risk analysis	LDA, GARCH Model	Statistical Analysis	Financial and Non-Financial Sectors	Measured cyber risk using LDA and GARCH model.
Tudosi et al., 2023)	Security Weakness in Firewalls	Vulnerabilities, penetration testing	Penetration Testing	Security Audit	Distributed Firewall Systems	Researched security weaknesses in distributed firewalls.
Darem et al., 2023)	Cyber Threats in Banking	Threat types, countermeasures	Literature Review	Literature Review	Banking and Financial Sector	Classified cyber threats and countermeasures in banking.

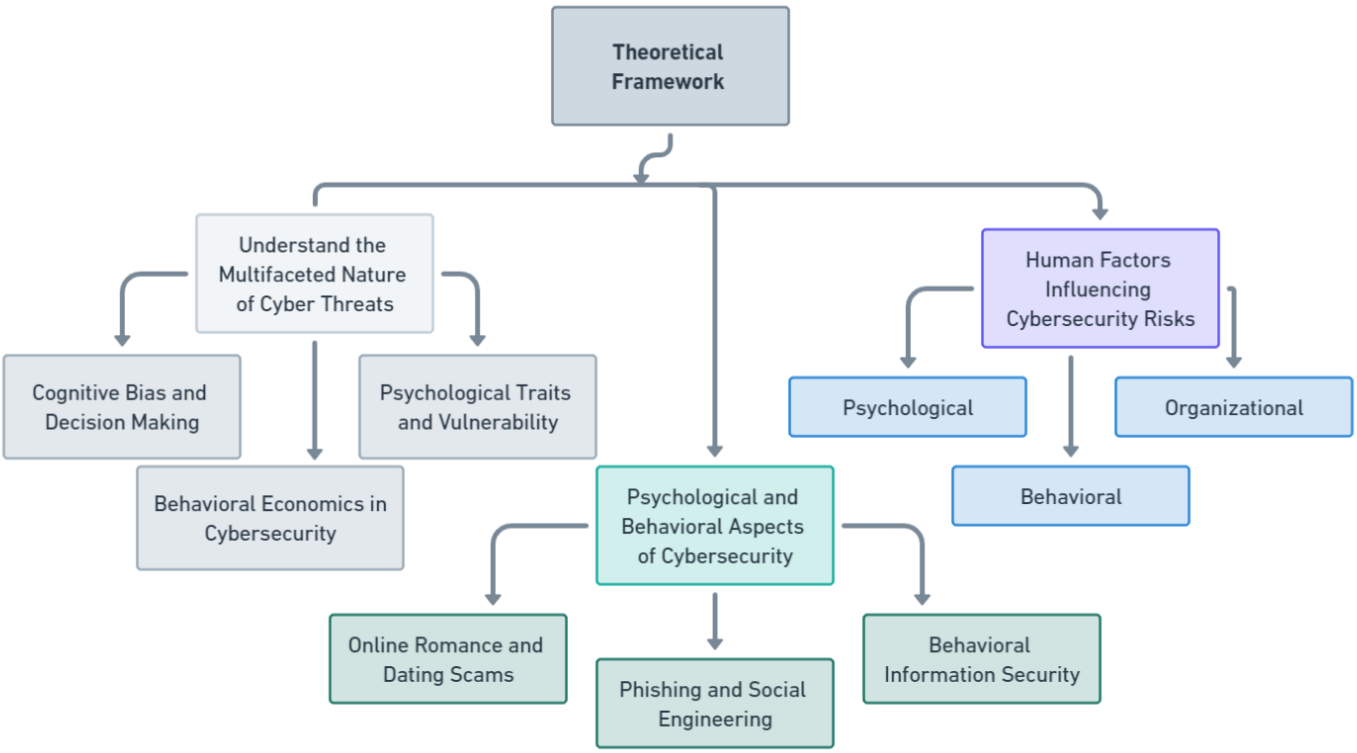


Figure 1. Theoretical framework

III. Theoretical Framework

The suggested theoretical framework on cybersecurity visually organizes the complex interplay of psychological, behavioral, and organizational factors that influence cybersecurity risks. It begins with an understanding the multifaceted nature of cyber threats, emphasizing the critical intersection of psychology and cybersecurity. This part deals with how cognitive biases, behavioral patterns, and psychological traits shape individual and organizational susceptibility to cyber-attacks, highlighting the importance of cybersecurity awareness and the need for behavior change. Key dimensions explored include cognitive bias and decision-making, where cyber threats exploit common biases; behavioral economics in cybersecurity, examining the gap between knowledge and practice; and psychological traits and vulnerability, focusing on how certain characteristics predispose individuals to cyber risks. **Error! Reference source not found.** f urther outlines the psychological and behavioral aspects of cybersecurity, detailing the strategies to mitigate vulnerabilities through understanding online scams, phishing, and the discrepancy between what users know and how they act. This leads to a categorization of human factors based on psychological, behavioral, and organizational dimensions, offering a comprehensive overview of how these elements contribute to the cybersecurity risk landscape.

1- Understand the multifaceted nature of cyber threats.

The intersection of psychology and cybersecurity is pivotal in unraveling the complexities of human vulnerabilities that cyber threats exploit. This exploration delves into the ways psychological traits, behavioral patterns, and cognitive biases shape individual susceptibility to cyber-attacks. By examining the mechanisms behind social engineering tactics, such as phishing and online scams, we underscore the critical role of cybersecurity awareness and the necessity for behavior change. This discussion is structured around three key dimensions:

Cognitive Bias and Decision Making: Cyber threats frequently leverage cognitive biases, like the inclination to trust familiar requests or downplay personal risk. The Elaboration Likelihood Model (ELM) provides a lens through which we can understand how individuals process and react to cybersecurity threats, emphasizing the role of cognitive biases in influencing susceptibility. This insight is crucial for designing interventions that

effectively counteract these biases, thereby enhancing cybersecurity measures.

- **Behavioral Economics in Cybersecurity**: Incorporating behavioral economics, such as understanding loss aversion and overconfidence bias, offers valuable perspectives on the challenges individuals and organizations face in adopting robust cybersecurity behaviors. The investigation into phishing and social engineering by Workman (2008), alongside Crossler et al. (2013)'s insights into behavioral information security, illustrates the critical gap between knowledge and practice. These findings suggest a compelling need for cybersecurity training that not only addresses habitual behaviors but also directly confronts cognitive biases impeding behavior change.
- **Psychological Traits and Vulnerability**: Certain psychological characteristics, such as a high degree of openness or elevated levels of trust, may predispose individuals to riskier online behaviors. Foundational research by Whitty & Buchanan (2012) and Buchanan & Whitty (2014) into online romance scams reveals how emotional vulnerabilities are targeted by cybercriminals. These studies highlight the importance of tailored cybersecurity education and awareness programs that account for the diverse psychological profiles of internet users.

2- *Psychological and Behavioral Aspects of Cybersecurity*

Focusing on how human factors impact cybersecurity risks, we engage in a detailed analysis of the strategies devised to mitigate these vulnerabilities. This comprehensive approach encompasses:

- **Online Romance and Dating Scams:** The work of Whitty & Buchanan (2012) and Buchanan & Whitty (2014) delves into the exploitation of psychological traits by cybercriminals, underlining the need for preventive measures that consider the victim's emotional and psychological state.
- **Phishing and Social Engineering**: Highlighted by Workman's (2008) research, the susceptibility to phishing attacks is often rooted in habitual online behaviors, suggesting a shift towards **behavior** change as a cornerstone of cybersecurity training.
- **Behavioral Information Security**: Crossler et al. (2013) emphasize the discrepancy between what internet users know and how they act, pointing towards an essential focus on bridging this gap through targeted behavioral interventions.

This multidimensional framework not only illuminates the multifaceted nature of cyber threats but also guides the development of nuanced strategies to bolster cybersecurity. By integrating insights from psychology, behavioral science, and cybersecurity research, we can craft more effective education and awareness programs, tailored interventions, and robust security measures that consider the complex interplay of human factors in the digital realm. To summarize the human factors that influence cybersecurity risks, we'll categorize these factors in Table 2 based on psychological, behavioral, and organizational dimensions. This approach helps in understanding how various human elements contribute to the cybersecurity risk landscape.

Table 2. The human factors that influence cybersecurity risks

Human Factor	Category	Short Description
Cognitive Biases	Psychological	Cognitive biases like overconfidence, confirmation bias, and availability heuristic can lead individuals to underestimate cybersecurity risks or ignore security warnings.
Lack of Awareness	Behavioral	Insufficient knowledge about cyber threats and safe online practices leads to risky behaviors, such as clicking on phishing links or using weak passwords.
Habitual Behavior	Behavioral	Routine actions performed without conscious thought, such as automatically opening email attachments, increase vulnerability to cyber threats.
Emotional Vulnerabilities	Psychological	Emotions such as fear, curiosity, or urgency can be exploited by cyber attackers to manipulate individuals into divulging confidential information or making hasty decisions.
Resistance to Change	Organizational	Individuals or organizations resistant to updating cybersecurity practices or technologies may maintain outdated defenses, making them more susceptible to new or evolving threats.
Social Influence	Psychological	Social norms and peer behaviors can influence an individual's cybersecurity practices, sometimes leading to riskier online activities if those around them engage in unsafe behaviors.
Psychological Safety	Organizational	A lack of psychological safety in organizations may discourage employees from reporting potential security threats or admitting to security mistakes, hindering effective threat management.
Decision Fatigue	Psychological	Repeated decision-making or constant alerts can lead to decision fatigue, reducing the quality of decisions over time and potentially leading to security oversights.

Security Usability Trade-offs	Behavioral	The perceived inconvenience of security measures can lead users to bypass or weaken these measures for the sake of usability or efficiency, compromising security.
Organizational Culture	Organizational	The overall culture of an organization, including its values, norms, and practices around cybersecurity, significantly influences the cybersecurity behaviors of its members.
Training and Education	Organizational	The quality and frequency of cybersecurity training and education directly impact an individual's ability to recognize and respond to cyber threats effectively.
Personal Accountability	Behavioral	An individual's sense of responsibility and accountability for maintaining cybersecurity practices affects their diligence in following security protocols.

This table highlights the multifaceted nature of human factors in cybersecurity risks, underscoring the need for comprehensive approaches that address psychological, behavioral, and organizational dimensions to enhance cybersecurity resilience.

3- CYBERSECURITY AWARENESS AND IMPACT

Cybersecurity awareness and its impact on mitigating cyber risks have become pivotal in the digital age. As cyber threats evolve in complexity and sophistication, the human element of cybersecurity—ranging from individual behavior to organizational culture—plays a crucial role in the effectiveness of security measures. This section aims to shed light on the critical aspects of cybersecurity awareness and its consequential impact on individuals and organizations, underlining the need for comprehensive awareness programs and the assessment of their efficacy. Parsons et al. (2014)'s examination of employee cybersecurity awareness levels across sectors emphasizes the gap between awareness and behavior, pointing to the need for engaging and continuous education programs that resonate with employees' daily practices.

Button et al. (2014) focus on the personal and familial impact of fraud, highlighting the emotional and psychological toll of cyber incidents. This underscores the importance of incorporating emotional support and counseling into post-incident response plans.

a. The Importance of Cybersecurity Awareness

1. **Foundational Awareness and Behavioral Change**: Cybersecurity awareness is not merely about disseminating information; it's about fostering a fundamental understanding and instigating behavioral change among users. Initiatives must move beyond generic advice to provide actionable, context-specific guidance tailored to diverse user groups.
2. **Organizational Culture and Cybersecurity Hygiene**: The role of organizational culture in cybersecurity cannot be overstated. A culture that prioritizes cybersecurity hygiene and encourages open communication about cyber risks can significantly enhance an organization's resilience to cyber threats.
3. **Psychological Safety and Reporting**: Creating an environment of psychological safety, where employees feel comfortable reporting potential threats without fear of reprimand, is crucial for early detection and response to cyber incidents.

b. Impact of Fraud

Enhancing cybersecurity awareness and understanding its impact is a complex, multifaceted endeavor that requires a holistic approach. By integrating insights from psychology, organizational behavior, and cybersecurity, we can develop more effective strategies for promoting cybersecurity hygiene and resilience. Future research and practice must focus on creating an informed, vigilant, and resilient digital society capable of defending against and recovering from cyber threats. Button et al. (2014) analyze the personal and familial impact of fraud through qualitative analysis, indicating the profound effects of online scams beyond financial loss. The Impact of Fraud can affect the following areas:

1. **Economic and Psychological Consequences**: The impact of cyber incidents extends beyond immediate financial loss, affecting the psychological well-being of victims and the reputation of organizations. Comprehensive risk management strategies must address these broader implications.
2. **Resilience and Recovery**: Building resilience against cyber threats involves not only preventative measures but also effective recovery plans that minimize the impact of breaches. This includes technical response mechanisms as well as support for affected individuals.
3. **Measuring the Effectiveness of Awareness Programs**: To truly gauge the impact of cybersecurity awareness initiatives, organizations must employ metrics that measure changes in behavior and culture, not just the dissemination of information. This could involve regular

simulations, phishing tests, and feedback mechanisms to assess and refine the programs continuously.

4- ECONOMIC AND SYSTEMIC PERSPECTIVES

The economic and systemic dimensions of cybersecurity underscore the significance of understanding cybercrime's financial impact and the strategic deployment of cybersecurity measures. This perspective involves analyzing the costs associated with cyber incidents, the economic rationale behind investments in cybersecurity, and the systemic integration of advanced technologies to fortify digital infrastructures. By exploring these aspects, we can better appreciate the economic drivers of cybercrime and the systemic strategies required to mitigate these threats effectively.

- **Economic Implications of Cybercrime**

1. **Direct and Indirect Costs**: Cyber incidents incur direct costs such as financial losses from theft, data breach response efforts, and legal expenditures. Indirect costs, including reputational damage, loss of customer trust, and long-term competitive disadvantage, can surpass direct costs and impact organizations for years.
2. **Economics of Cybersecurity Investments**: Investing in cybersecurity is often viewed through a cost-benefit lens, where the potential losses from cyber incidents are weighed against the costs of implementing security measures. Decision-making models, such as Return on Security Investment (ROSI), can help organizations optimize their cybersecurity spending.
3. **Cyber Insurance as a Risk Management Tool**: The growth of the cyber insurance market reflects an economic approach to managing cyber risks, transferring some of the financial risks to insurers. However, the effectiveness and coverage of cyber insurance policies remain areas for further research and development.

- **Systemic Approaches to Cybersecurity**

1. **Integration of Advanced Technologies**: The development and application of technologies such as Artificial Intelligence (AI), Machine Learning (ML), and blockchain in cybersecurity offer new avenues for detecting and mitigating cyber threats. For instance, AI and ML can enhance anomaly detection in network traffic, while blockchain offers potential for secure, tamper-proof systems.
2. **Economic Analysis of Online Crime (Moore et al., 2019)**: This study provides an insightful theoretical exploration of the economics behind online crime, discussing the motivations of cybercriminals and the financial strategies for cyber defense. It lays the groundwork for understanding the economic incentives that drive cybercrime and the allocation of resources for cybersecurity.
3. **Cybersecurity System Development (Hephzipah et al., 2023)**: The creation of systems for anomaly detection using Artificial Neural Networks exemplifies the systemic integration of advanced technologies in cybersecurity efforts. These innovations highlight the shift towards automated and intelligent security solutions.

- **Bridging Economic and Systemic Insights**

1. **Cost-Effective Security Solutions**: The economic perspective encourages the development of cost-effective cybersecurity solutions that balance financial investment with security efficacy. This includes evaluating the cost savings of automated security systems versus traditional approaches.
2. **Public-Private Partnerships**: Collaborations between governments and the private sector can leverage economic and systemic strengths to enhance national and global cybersecurity postures. These partnerships facilitate the sharing of threat intelligence, economic resources, and technological innovations.
3. **Future Research Directions**: Future research should focus on quantifying the effectiveness of advanced cybersecurity technologies in economic terms, developing frameworks for strategic cybersecurity investment, and exploring the impact of regulatory environments on the economic aspects of cybersecurity.

The economic and systemic perspectives on cybersecurity provide critical insights into the financial impact of cybercrime and the strategic implementation of advanced technologies to combat these threats. Understanding the economic drivers behind cybercrime and investing in systemic cybersecurity solutions is essential for organizations seeking to navigate the complex cyber threat landscape effectively. Future endeavors in this field should aim to bridge economic analysis with technological innovation, fostering a security ecosystem that is both economically viable and resilient against emerging threats.

To address the human factors influencing cybersecurity risks, a range of strategies can be employed. These strategies are designed to mitigate vulnerabilities by enhancing awareness, changing behaviors, and cultivating a security-oriented organizational culture. The following table outlines these strategies, categorized by the human factors they aim to mitigate.

Table 3. Strategies to mitigate the vulnerabilities of human factors influencing cybersecurity risks.

Human Factor	Strategy	Short Description
Cognitive Biases	Behavioral Interventions & Nudges	Use psychological interventions and nudges to counteract biases, such as implementing clear, concise security warnings that consider user psychology to encourage safer behaviors.
Lack of Awareness	Comprehensive Education & Training	Provide ongoing, engaging cybersecurity education and training programs that cover the latest threats and safe practices, tailored to different roles within an organization.
Habitual Behavior	Security Automation & Simplification	Implement security measures that automate safe practices or simplify security decisions, reducing reliance on habitual behaviors that may lead to vulnerabilities.
Emotional Vulnerabilities	Emotional Intelligence Training	Enhance emotional intelligence to help individuals recognize and manage emotions that could be exploited by cyber threats, such as training on recognizing phishing attempts that use urgency or fear.
Resistance to Change	Change Management & Incentivization	Employ change management strategies to gradually introduce new cybersecurity practices, coupled with incentives for adoption, to overcome resistance.
Social Influence	Peer-led Initiatives & Social Proof	Leverage peer influence by showcasing positive security behaviors and outcomes through peer-led initiatives and highlighting widespread adoption of security practices (social proof) to encourage conformity to secure behaviors.
Psychological Safety	Open Communication & Non-punitive Reporting Policies	Foster an organizational culture that encourages open communication about cybersecurity issues and implements non-punitive reporting policies to ensure employees feel safe reporting threats and mistakes.
Decision Fatigue	Alert Prioritization & Simplification	Reduce decision fatigue by prioritizing and simplifying security alerts and decisions, ensuring that individuals deal with fewer, more meaningful warnings and choices.
Security Usability Trade-offs	User-Centric Security Design	Design security measures with a focus on usability, ensuring that security practices do not significantly hinder user experience, thereby reducing the temptation to bypass security measures.
Organizational Culture	Culture Building & Leadership Engagement	Cultivate a security-first organizational culture through leadership engagement, where top management demonstrates a commitment to cybersecurity and sets the tone for the organization.
Training and Education	Tailored Training Programs & Continuous Learning	Develop tailored training programs that address the specific needs and vulnerabilities of different user groups within an organization, coupled with continuous learning opportunities to keep pace with evolving cyber threats.
Personal Accountability	Accountability Measures & Personal Incentives	Implement measures that hold individuals accountable for their cybersecurity behaviors, coupled with personal incentives for adhering to security protocols, to encourage personal responsibility.

These strategies represent a holistic approach to mitigating cybersecurity vulnerabilities associated with human factors. By addressing the root causes of these vulnerabilities, organizations can enhance their overall cybersecurity posture and resilience against threats.

IV. DISCUSSION

The susceptibility to cyber scams is a complex phenomenon influenced by a confluence of psychological, behavioral, technical, and environmental factors. This discussion synthesizes insights from various studies to understand how these

dimensions interact and influence an individual's likelihood of falling victim to cyber scams. Psychological factors play a pivotal role in susceptibility to cyber scams. Whitty and Buchanan's (2012) exploration into the psychological manipulation used in online romance scams reveals how scammers exploit emotional vulnerabilities. This is further supported by Buchanan and Whitty's (2014) study, which identifies loneliness and risk-taking behavior as psychological characteristics making individuals more susceptible to these scams. The role of cognitive biases, such as the optimism bias, which leads individuals to underestimate their risk of becoming scam victims, is a crucial aspect of this vulnerability. Behavioral aspects, including routine activities and security fatigue, significantly influence susceptibility to cyber scams. Workman's (2008) discussion on habitual behaviors and complacency highlights how routine online activities can increase exposure to cyber scams. Crossler et al. (2013) extend this understanding by linking frequent online activities with increased scam exposure. The need for behavioral change, emphasizing cyber hygiene practices, is evident in combating these threats.

The technical dimension of cyber scam susceptibility is nuanced. While a lack of technical understanding leaves individuals vulnerable to sophisticated scams, overconfidence in one's technical abilities can also lead to underestimating scammer sophistication. Parsons et al. (2014) emphasize the importance of continuous cybersecurity education to bridge this knowledge gap. The evolving nature of cyber threats necessitates keeping abreast of the latest security measures and understanding the technicalities of scams. The environmental dimension, encompassing social and institutional factors, shapes individuals' susceptibility to cyber scams. Button et al. (2014) highlight how social networks and cultural norms influence perceptions and responses to cyber scams. The role of institutional policies in creating safer cyber environments is critical. This includes not only regulations and laws but also organizational cultures that prioritize cybersecurity awareness. An interdisciplinary approach is vital in addressing the multifaceted nature of cyber scams. Moore et al. (2019) provide insights into how combining psychology, behavioral science, information technology, and social sciences can lead to more effective scam prevention strategies. Future research should focus on developing integrated models that consider all these dimensions to devise comprehensive prevention and intervention strategies. Future research in cyber scams should aim at developing more holistic models that incorporate psychological, behavioral, technical, and environmental factors. The development of predictive models using machine learning to identify potential scam victims based on these dimensions could be a significant step forward. Additionally, there is a need for more empirical research to test the effectiveness of different prevention and intervention strategies across various demographic groups.

The susceptibility to cyber scams spanning the psychological and behavioral aspects of cybersecurity, awareness and impact, economic and systemic perspectives, and strategies to mitigate vulnerabilities—reveals a complex interplay between human factors and cybersecurity risks. This comprehensive exploration underscores the importance of adopting a multidimensional approach to cybersecurity, one that goes beyond technical measures to include psychological insights, behavioral changes, and organizational culture shifts. Our exploration began with an in-depth look at how psychological traits and behavioral patterns influence individuals' susceptibility to cyber threats. Studies like those by Whitty & Buchanan and Workman highlight the critical role of emotional vulnerabilities and habitual behaviors in cybersecurity breaches. These insights suggest that effective cybersecurity measures must account for human psychology, emphasizing the need for educational programs that not only inform but also engage users emotionally and cognitively to foster a deeper understanding and change in behavior. The discussion on cybersecurity awareness and its impact stressed the pivotal role of knowledge and organizational culture in mitigating cyber risks. Awareness programs, as indicated by research from Parsons et al., must transcend basic information dissemination to instill a genuine comprehension of cyber threats and foster a proactive cybersecurity posture among individuals and within organizations. The emotional and psychological impact of cyber incidents, highlighted by Button et al., further reinforces the need for comprehensive support systems that address the wide-ranging consequences of cyber breaches. The economic and systemic perspectives on cybersecurity introduced a broader view of the challenges and strategies in combating online crime. The discussion covered the economic analysis of online crime by Moore et al. and the development of cybersecurity systems, showcasing the multifaceted nature of cybersecurity efforts that include not only prevention and detection but also a thorough understanding of the economic incentives behind cyber-attacks. Addressing the human factors that influence cybersecurity risks necessitates targeted strategies that encompass educational, organizational, and technological interventions. The proposed strategies aim to counteract cognitive biases, enhance

awareness and training, promote a security-centric organizational culture, and implement user-centric security designs. These approaches are designed to build resilience by not only improving security practices but also by fostering an environment where cybersecurity is a shared responsibility.

This comprehensive analysis illustrates that while technological advancements are crucial in combating cyber threats, understanding, and influencing human behavior and organizational culture are equally important. The interrelation between human factors and cybersecurity underscores the need for a holistic approach that integrates technical solutions with psychological and behavioral insights. Future cybersecurity efforts should focus on developing adaptive, user-friendly security measures, promoting continuous education and awareness, and fostering a culture of security that empowers individuals to act as the first line of defense against cyber threats. The fight against cyber threats is not just a technical challenge but a human one as well. The strategies and insights discussed throughout these topics highlight the importance of a multifaceted approach that addresses the complex nature of cybersecurity. By focusing on the human elements of cybersecurity, organizations can enhance their resilience against an ever-evolving threat landscape, ensuring a safer digital environment for all users.

V. LIMITATIONS AND CHALLENGES

One of the primary challenges in understanding cyber scam susceptibility is the complexity of psychological profiling. While studies like those by Whitty and Buchanan (2012) have shed light on the psychological traits that make individuals vulnerable to scams, the diversity and complexity of human psychology make it difficult to create a one-size-fits-all profile. Psychological factors such as trust, fear, and loneliness are not uniformly distributed across populations, and their influence on scam susceptibility can vary greatly depending on individual circumstances and experiences. Another significant challenge is the predictability and variability of human behavior. As Workman (2008) notes, habitual behaviors and security fatigue can lead to increased vulnerability to cyber scams. However, predicting which behaviors will lead to susceptibility is complex, as they can be influenced by a wide range of factors, including personal habits, cultural background, and even current emotional states. This variability makes it challenging to develop universally effective behavioral interventions.

The rapid evolution of technology and the sophistication of cyber scams present another major challenge. Technical knowledge, as discussed by Parsons et al. (2014), is crucial in recognizing and avoiding scams. However, as cyber scams become more sophisticated, keeping up with the necessary technical knowledge becomes increasingly difficult for the average user. This gap leaves even technically savvy individuals vulnerable to new and evolving scam tactics. The influence of environmental and cultural factors on scam susceptibility is a complex area with significant limitations in current research. Button et al. (2014) highlight the role of social networks and cultural norms in shaping responses to cyber scams. However, the vast diversity in cultural and social environments across different regions and communities makes it challenging to develop universal guidelines or preventive measures that are effective in all contexts. The interdisciplinary nature of cyber scam research, involving psychology, behavioral science, information technology, and social sciences, presents its own set of challenges. Integrating insights from these diverse fields, as Moore et al. (2019) suggest, is crucial but also complex. Different disciplines have different methodologies, terminologies, and focus areas, which can make interdisciplinary research challenging. Given these limitations and challenges, future research in the field of cyber scams needs to focus on developing more nuanced and individualized approaches. This includes creating more sophisticated psychological profiles, understanding the variability in behavioral responses, keeping pace with evolving technical threats, and considering the diverse environmental and cultural contexts in which scams occur.

VI. OPEN PROBLEMS

These open problems highlight the dynamic and multifaceted nature of research in cyber scam susceptibility. Addressing these challenges requires ongoing, collaborative efforts from researchers, practitioners, and policymakers.

Evolving Nature of Cyber Scams: One of the most significant open problems in the field of cyber scam susceptibility is the continuously evolving nature of cyber scams themselves. As technology advances, scammers develop new and more sophisticated methods to exploit users. This constant evolution presents a moving target for researchers and cybersecurity professionals, making it challenging to develop long-term, effective countermeasures. Understanding the latest trends in scam tactics and

developing predictive models that can adapt to these changes remains a critical, unresolved challenge.

Psychological Profiling and Predictive Analysis: Another open problem is the development of accurate psychological profiles that can predict an individual's susceptibility to cyber scams. Current research, such as the work by Whitty and Buchanan (2012), provides valuable insights into common psychological traits of scam victims. However, creating comprehensive profiles that consider the wide range of human emotions, behaviors, and experiences is an ongoing challenge. Moreover, ethical considerations in using such profiles for predictive analysis need to be addressed, ensuring privacy and fairness.

Behavioral Change and User Education: Despite the recognition of the importance of user behavior in cybersecurity, effectively changing this behavior remains a complex issue. Educational and awareness programs have had varying degrees of success, as noted by Crossler et al. (2013). Developing more effective methods to alter user behavior, particularly in ways that are sustainable and adaptable to different demographic groups, is an open problem in the field.

Technical Solutions vs. Human Factors: The balance between technical solutions and human factors in preventing cyber scams is an area of ongoing debate and research. While technical measures are essential, they often fail to address the human element of cybersecurity. As Parsons et al. (2014) suggest, increasing technical security measures does not always equate to better protection if users are not aware or do not understand how to use these measures effectively. Finding the right balance and integration of technical and human-centric approaches remains a significant challenge.

Cultural and Environmental Influences: The impact of cultural and environmental factors on scam susceptibility, as discussed by Button et al. (2014), is an area that requires further exploration. Cultural norms and values significantly influence online behavior and responses to scams, yet there is a lack of comprehensive research that spans different cultural contexts. Understanding these influences in a globalized online environment is crucial for developing effective, culturally sensitive prevention strategies.

Interdisciplinary Collaboration: Finally, the need for interdisciplinary collaboration in addressing cyber scam susceptibility is an area with great potential but also significant challenges. Bringing together expertise from psychology, information technology, social sciences, and cybersecurity, as Moore et al. (2019) advocate, is essential for a holistic understanding of cyber scams. However, fostering effective collaboration across these diverse fields, each with its methodologies and perspectives, remains a complex and unresolved issue.

VII. CONCLUSIONS

The study of cyber scam susceptibility reveals a complex landscape where psychological, behavioral, technical, and environmental factors intertwine. Our analysis underscores that no single dimension can fully explain why individuals fall victim to cyber scams. Instead, it is the interplay of these factors that shapes susceptibility. Psychologically, individuals' cognitive biases and emotional states significantly influence their vulnerability. Behaviorally, routine online activities and a lack of cyber hygiene practices contribute to increased risk. Technically, the rapid evolution of scamming techniques often outpaces the average user's knowledge and preparedness. Environmentally, cultural and social contexts play a crucial role in shaping individuals' awareness and responses to scams. This multifaceted nature of susceptibility presents significant challenges in developing effective prevention and intervention strategies. It is clear that efforts to combat cyber scams must be as dynamic and multifaceted as the scams themselves. This involves not only educating the public about common scamming tactics but also fostering a deeper understanding of the psychological and behavioral aspects that underlie scam susceptibility. Future research should focus on developing more nuanced models that consider these diverse factors, aiming to create targeted interventions. Additionally, there is a need for more empirical research to test the effectiveness of different prevention strategies across various demographic groups. In conclusion, understanding and mitigating the risk of cyber scams requires a concerted effort that spans multiple disciplines and perspectives. By acknowledging and addressing the complex interplay of factors that contribute to scam susceptibility, we can develop more effective strategies to protect individuals in the digital age.

REFERENCES

[1] Whitty, M. T., & Buchanan, T. (2012). The online romance scam: A serious cybercrime. *Cyberpsychology, Behavior, and Social Networking*, 15(3), 181-183.

[2] Buchanan, T., & Whitty, M. T. (2014). The online dating romance scam: Causes and consequences of victimhood. *Psychology, Crime & Law*, 20(3), 261-283.

[3] Workman, M. (2008). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American Society for Information Science and Technology*, 59(4), 662-674.

[4] Crossler, R. E., Johnston, A. C., Lowry, P. B., Hu, Q., Warkentin, M., & Baskerville, R. (2013). Future directions for behavioral information security research. *Computers & Security*, 32, 90-101.

[5] Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., & Jerram, C. (2014). Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q). *Computers & Security,* 42, 165-176.

[6] Button, M., Lewis, C., & Tapley, J. (2014). Not a victimless crime: The impact of fraud on individual victims and their families. *Security Journal*, 27(1), 36-54.

[7] Moore, T., Clayton, R., & Anderson, R. (2019). The economics of online crime. *Journal of Economic Perspectives,* 23(3), 3-20.

[8] Hephzipah, J. J., Vallem, R. R., Sheela, M. S., & Dhanalakshmi, G. (2023). An efficient cyber security system based on flow-based anomaly detection using Artificial neural network. *Mesopotamian Journal of Cybersecurity*, *2023*, 48-56.

[9] Marican, M., Razak, S., Selamat, A., & Othman, S. (2023). Cyber security maturity assessment framework for technology startups: a systematic literature review. Ieee Access, 11, 5442-5452. https://doi.org/10.1109/access.2022.3229766

[10] Shang, Y., Wang, K., Tian, Y., Zhou, Y., Ma, B., & Liu, S. (2023). Theoretical basis and occurrence of internet fraud victimisation: based on two systems in decision-making and reasoning. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1087463

[11] Srivastava, A., Mukherjee, S., Datta, B., & Shankar, A. (2023). Impact of perceived value on the online purchase intention of base of the pyramid consumers. International Journal of Consumer Studies, 47(4), 1291-1314. https://doi.org/10.1111/ijcs.12907

[12] Angelelli, M., Arima, S., Catalano, C., & Ciavolino, E. (2023). Cyber-risk Perception and Prioritization for Decision-Making and Threat Intelligence. *ArXiv*, abs/2302.08348. https://doi.org/10.48550/arXiv.2302.08348.

[13] Ashwini, S., Sinha, M., & Sabarinathan, C. (2023). Implementation of Intrusion Detection Model for Detecting Cyberattacks Using Support Vector Machine. *Advances in Science and Technology*, 124, 772 - 781. https://doi.org/10.4028/p-6nyqo1.

[14] Kim, S., & Song, S. (2023). Cyber risk measurement via loss distribution approach and GARCH model. *Communications for Statistical Applications and Methods*. https://doi.org/10.29220/csam.2023.30.1.075.

[15] Tudosi, A., Graur, A., Balan, D., & Potorac, A. (2023). Research on Security Weakness Using Penetration Testing in a Distributed Firewall. *Sensors*, 23 5. https://doi.org/10.3390/s23052683.

[16] A. A. Darem, A. A. Alhashmi, T. M. Alkhaldi, A. M. Alashjaee, S. M. Alanazi and S. A. Ebad, "Cyber Threats Classifications and Countermeasures in Banking and Financial Sector," in IEEE Access, vol. 11, pp. 125138-125158, 2023, doi: 10.1109/ACCESS.2023.3327016.

[17] Eze, O., Okpa, J., Onyejegbu, C., & Ajah, B. (2023). Cybercrime: victims' shock absorption mechanisms. https://doi.org/10.5772/intechopen.106818

[18] Khaldi, L., Elabed, A., & Khanchoufi, A. (2023). Quantitative assessment of the relative impacts of different factors on flood susceptibility modelling: case study of fez-meknes region in morocco. E3s Web of Conferences, 364, 02005. https://doi.org/10.1051/e3sconf/202336402005

[19] Watkins, M. (2023). Designing an effective organizational culture to guard against the cyber risks of emerging technologies. Journal of Healthcare Management, 68(4), 239-250. https://doi.org/10.1097/jhm-d-23-00097

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

Artificial Intelligence in Education Predicting College Plans of High School Students

*Aws I. AbuEid [1], *, Radhia Zaghdoud[2], Olfa Ben Rhaiem[2] , Marwa Amara [2],Achraf Ben Miled[2,3] , Ashraf F. A. Mahmoud[2], Faroug A. Abdalla[2], Chams Jabnoun[2], Aida Dhibi[2], Ahlem Fatnassi[2] , Firas M. Allan[2], Mohammed Ahmed Elhossiny[4,5], , Imen Ben Mohamed[2], Marwa Anwar Ibrahim Elghazawy[4], Majid A. Nawaz[2], Salem Belhaj[2]*

[1]*Faculty of Computing Studies, Arab Open University,Amman, Jordan*

[2]*Computer Science Department, Science College, Northern Border University, Arar, Kingdom of Saudi Arabia*

[3] *Artificial Intelligence and Data Engineering Laboratory, LR21ES23, Faculty of Sciences of Bizerte, University of Carthage, Tunisia*

[4]*Applied College, Northern Border University, Arar, Saudi Arabia*

[5]*Faculty of Specific Education, Mansoura University, Mansoura, Egypt.*

**Corresponding Author: Email: a_abueid@aou.edu.jo*

ABSTRACT

The study introduces AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits), a predictive model utilizing artificial intelligence to forecast high school students' college plans. It achieves promising results with an accuracy rate of 84.88% through advanced techniques like hyperparameter tuning using scikit-learn's GridSearchCV. The model's development process, including data preprocessing and feature engineering, is detailed. Results show improved accuracy, precision, recall, and F1 scores, particularly for students with college plans. Visualizations aid in interpreting outcomes, supporting stakeholders in educational decision-making. The AIRPCP model has significant implications for educators, policymakers, and researchers, offering insights to inform educational planning and policy development, ultimately supporting students' aspirations.

Keywords: Artificial Intelligence, machine learning, python libraries, AIRPCP

1. INTRODUCTION

Pursuing a university education is a pivotal decision in the lives of high school students, significantly shaping their prospects and career paths. This critical choice carries immense weight for educational institutions and policymakers alike, as it forms the basis for planning admissions procedures and anticipating workforce requirements, thus informing management decision-making support strategies. Yet, obtaining precise insights into students' college aspirations presents a formidable challenge.

This article introduces a novel approach, the Artificial Intelligence in Education Predicting College Plans (AIEPCP) model, which leverages advanced artificial algorithms, particularly neural networks, to explore high school students' intentions regarding college attendance. These intentions are often considered highly personal and sensitive information, making students hesitant to share them openly [1]. This research aims to develop a new predictive model AIEPCP capable of estimating high school students' inclinations toward college

attendance. By doing so, it serves as a valuable resource for informing public universities and labor departments about the expected influx of new students or potential job seekers.

Furthermore, this study embarks on a journey into the domain of predictive analysis within the education sector, specifically focusing on forecasting high school students' college intentions. Employing the potent capabilities of Artificial Intelligence (AI), it harnesses the abundance of available student data. The ultimate objective is to discern the key factors differentiating high school students with a propensity for higher education from those who do not lean in that direction.

The motivation behind this research is rooted in the pressing need to address the challenges surrounding precise college enrollment predictions. These predictions are fundamental to educational planning and critical in anticipating labor market trends and optimizing resource allocation. Acknowledging the influential role of parents and the critical nature of factors such as gender, IQ, income, and parental encouragement [2], this study aspires to enhance decision-making processes within education.

Kaggle, a renowned platform for data science competitions, provided the dataset used in this study and datasets [3] sourced from government records. This dataset underwent rigorous curation and preprocessing to ensure data quality and privacy protection. Personal identifiers, such as student names, were thoughtfully desensitized to safeguard individual privacy and replaced with unique Student IDs.

The selected explanatory variables for analysis encompass gender, IQ, parental income, and parental encouragement to pursue higher education. According to experienced government officials, these variables have emerged as the most influential factors shaping high school students' college intentions [4]. Given the scale and intricacy of the dataset, this study offers access to a substantial sample consisting of up to 8,000 data samples. This sample size is sufficient for constructing a robust predictive model, particularly one based on neural networks, enabling comprehensive analysis.

This study is inherently concerned with a classification problem, where the objective is to categorize high school students into distinct groups based on their college plans, aligning seamlessly with the predictive capabilities of AI. Furthermore, this article endeavors to illuminate the path toward a more precise understanding of high school students' college aspirations through the innovative fusion of Artificial Intelligence and educational data analysis to provide valuable insights that will benefit educational institutions, policymakers, and students.

The main objective and hypothesis is to develop a predictive AIEPCP model leveraging neural networks to demonstrate high predictive accuracy in forecasting high school students' college intentions, outperforming traditional predictive methods.

2. LITERATURE REVIEW

The literature review section provides a comprehensive overview of prior research and studies relevant to the predictive analysis of high school students' college intentions and the use of artificial intelligence in this educational context.

The study by Pan et al. [5] delves into the intricate relationship between preschool cognitive and behavioral skills and their potential influence on indicators of college enrollment, focusing primarily on a sample of youth residing in low-income areas of Chicago, predominantly comprising Black and Hispanic students. The findings of this research contribute to the broader discourse on the early predictors of future educational attainment, especially within disadvantaged communities.

While the study reveals that most early cognitive and behavioral skills exhibit only weak to moderate associations with later college enrollment, a noteworthy standout is the role of preschool attention and impulsivity control. This particular skill emerges as a relatively strong predictor of college enrollment, sparking interest in the potential impact of early attentional and self-regulatory abilities on long-term educational trajectories.

Furthermore, it dispels concerns regarding early behavioral difficulties as substantive predictors of college enrollment, indicating that cognitive capabilities, particularly those associated with attention and executive functioning, play a more pivotal role in this context. The findings enrich our understanding of the multifaceted nature of educational attainment and underscore the need for comprehensive, context-aware approaches to educational intervention and policy development.

Another study conducted by Ye [6] offers a unique perspective on the dynamics of college choice behavior within centralized admission systems, shedding light on

the critical role of precise predictions in improving educational outcomes. The research underscores the complex interplay between students' strategic college choices and the outcomes of centralized admissions, ultimately highlighting the need for informed decision-making. In response to this phenomenon, the study implements a large-scale randomized experiment involving a substantial cohort of students (N=32,834). This experiment provides treated students with valuable resources: (a) an application guidebook or (b) a guidebook coupled with a school workshop. The results of this intervention underscore the significance of informing students about selecting colleges and majors based on precise predictions of admission probabilities.

The experiment outcomes indicate that offering guidance to students regarding college and major selection, grounded in accurate predictions of admission probabilities, can yield notable improvements in aligning students with colleges that match their academic abilities. Specifically, the study demonstrates an enhancement in the student–college academic fit by 0.1 to 0.2 standard deviations among those who complied with the intervention without significantly altering their college-major preferences. It illuminates the intricate relationship between students' strategic choices and the outcomes of these systems. Moreover, the study underscores the potential of targeted interventions to enhance college choice behaviors and academic outcomes for students navigating centralized admission processes.

The findings of this research bear implications for policymakers, educators, and administrators involved in designing and managing centralized admissions systems. They emphasize the importance of providing students with precise information and guidance to make informed college choices, ultimately ensuring a more equitable and effective educational landscape to underscore the significance of precise predictions in shaping educational outcomes.

In recent years, integrating Artificial Intelligence (AI) into educational analytics has marked a transformative shift in how institutions and researchers approach student outcomes, employability predictions, and educational decision-making. The study conducted by Yan and Chi [7] highlights the pivotal role of AI, specifically the decision tree classification algorithm, in predicting college students' employment outcomes based on their educational background and work experience. The study underscores the practical value of AI-driven analytics in higher vocational education. By utilizing the decision tree classification algorithm, the research identifies key determinants of students' employment success and designs prediction models that inform enrollment strategies, regional employment dynamics, and the types of employment opportunities available. This data-driven approach aligns with the broader educational trend of evidence-based decision-making. It can potentially guide employment guidance and talent development programs in higher vocational colleges. The study contributes to the body of knowledge that recognizes AI's transformative potential in providing data-driven insights for educational institutions and policymakers, ultimately fostering more informed and effective decision-making.

Online education has witnessed exponential growth in recent years, transforming the teaching and prediction of students' academic performance. The authors Jiao et al. [8] emphasize the pivotal role of artificial intelligence (AI) and data-driven learning models in offering fresh insights into students' learning behaviors and strategies. These models leverage educational data mining and learning analytics techniques to unlock the potential of extensive datasets, shedding light on how students learn and how their learning performance can be optimized. Researchers have employed various AI methods to construct prediction models, including evolutionary computation, deep learning, decision trees, and Bayesian networks. The study contributes significantly to the field of AI-enabled academic performance prediction. Furthermore, the research introduces an AI model, particularly genetic programming, designed to forecast students' academic performance accurately and offers analytical.

The study exemplifies the evolving landscape of AI-driven academic performance prediction in online education. It underscores the significance of data-driven learning prediction models, the challenges associated with AI algorithms, and the potential of evolutionary computation to address these challenges. This research aligns with the broader educational trend of evidence-based decision-making and data-driven improvements in learning outcomes.

Integrating data mining and artificial intelligence (AI) has brought about transformative changes in the educational sector, enabling educational institutions to harness the power of available data for informed decision-making. While data mining has long been recognized for its significance in the business world, its adoption in schools, universities, and colleges has become increasingly prevalent, with a particular focus on

improving educational policies and practices. The authers Gumba, etc. [9] explore a recent study that employs classification algorithms to predict student admission to Information Technology Education (ITE) programs, shedding light on the valuable insights it offers to educational policies and strategies, ultimately enhancing the quality of education offered [9]. The utilization of these techniques in education mirrors their application in the business world, emphasizing the importance of data-driven decision-making. The researchers employ four classification algorithms: Decision Tree, K-Nearest Neighbor, Logistic Regression, and Naive Bayes. This diverse set of algorithms underscores the multifaceted nature of predictive analysis, offering a comprehensive view of student admission considerations.

A crucial aspect of this study involves the selection of predictors that influence student admission decisions. Eight predictors were chosen based on correlation coefficient evaluation, with mathematical skill emerging as the strongest predictor, with a correlation coefficient of 0.767. This highlights the significance of mathematical proficiency in the context of ITE program admission.

In addition to predictor selection, the study evaluates the performance of each classifier using various metrics, including accuracy, precision, recall, and the F1 score. These metrics provide a robust assessment of the classifiers' predictive capabilities. Notably, the K-Nearest Neighbor classifier exhibited the highest accuracy, with a score of 93.18%, precision reached 97.98%, recall stood at 88.13%, and the F1 score achieved 92.76%. These metrics collectively demonstrate the effectiveness of the K-Nearest Neighbor algorithm in predicting student admission.

In summary, this study exemplifies the growing trend of leveraging data mining and AI techniques in the educational sector, specifically in the context of student admission to Information Technology Education programs. Selecting relevant predictors and evaluating classification algorithms' performance contribute to more informed admission decisions, ultimately benefiting educational institutions and aspiring students.

This study aims to address critical gaps in the existing literature by investigating the intersection of predictive analytics and educational decision-making, specifically focusing on high school students' college intentions. While previous research has explored related domains such as early predictors of educational attainment, college choice behavior, and the application of artificial intelligence (AI) in educational settings, a comprehensive investigation into the predictive analysis of high school students' college plans using advanced AI techniques remains absent.By focusing on this specific area, researchers can significantly contribute to the body of knowledge by unraveling the complex interplay of factors influencing students' decisions to pursue higher education. This includes not only cognitive and behavioral predictors, but also the critical role of external factors such as socioeconomic background, access to resources, and institutional support. Furthermore, the current literature exhibits a dearth of studies utilizing advanced AI methodologies such as hyperparameter optimization and ensemble learning approaches for predictive modeling in educational contexts.

Additionally, existing literature often falls short in providing comprehensive evaluations of predictive models' performance metrics, hindering our understanding of their effectiveness and generalizability.

Bridging these identified gaps in the literature will not only contribute to an advanced theoretical understanding but also offer practical solutions for improving educational planning, policy formulation, and student outcomes. This necessitates fostering interdisciplinary collaboration between experts in education, data science, and AI.

3. DATA COLLECTION AND PREPROCESSING

3.1 Description of the Dataset

The dataset employed in this investigation was sourced from Kaggle and comprises numerical and textual data to safeguard privacy and mitigate the risk of information disclosure; distinct Student IDs have been anonymized and substituted for the students' identities. The dataset originates from authentic enrollment records of the city for the preceding year, providing valuable insights into the college aspirations of high school students. The central focus of interest centers around the "Plan" column, which indicates whether a high school student intends to pursue a bachelor's education. In addition to the primary target variable, the dataset includes a variety of explanatory variables available for comprehensive analysis.

1. Student ID: is a distinct and individual identifier assigned to each student in the dataset. This identifier is unique to each student, ensuring that no two students

share the same Student ID. Student ID values span from 1 to 8000, encompassing all the students in the dataset and allowing for the unique identification of each student within this specified range. This unique identifier is a fundamental dataset component, enabling precise tracking and referencing of individual student records.

2. Gender: The gender of the student, with two categories: "male" and "female."

Figure 1 illustrates the distribution of high school students based on their gender. The chart visually represents the number of students falling into two gender categories: "female" and "male."

- Female (4126): This bar in the chart represents the count of female high school students, and the numeric value indicates that there are 4,126 female students in the dataset, 52%.

- Male (3874): The adjacent bar signifies the count of male high school students, and the associated numeric value indicates that there are 3,874 male students in the dataset,48%.

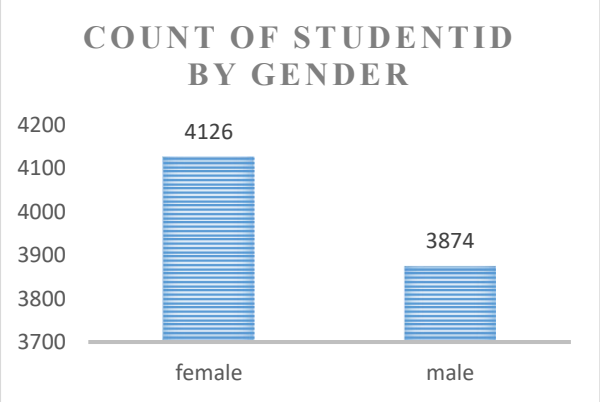


Figure 1: Count of Student ID by gender

Figure 1 lets us quickly grasp the gender distribution among high school students in the dataset, providing insights into the relative proportions of female and male students. This information is valuable for understanding the gender demographics of the sample and can be essential for further analysis and decision-making in various educational contexts.

3. Parent income: The parents' annual income, measured in US dollars, falls from $4,500 to $82,390.

Figure 2 provides an overview of the dataset's parental income distribution among high school students. The chart visually represents the spread of parental annual income in US dollars and highlights the range within which these incomes fall.

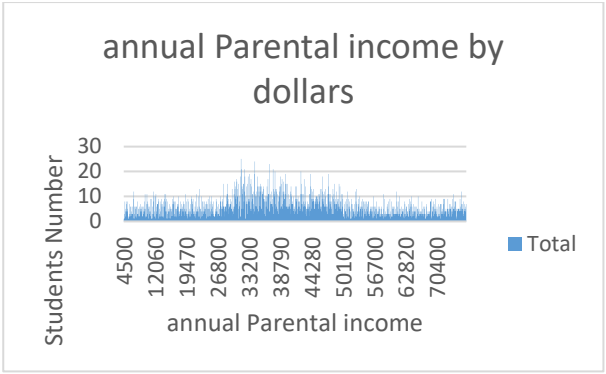


Figure 2: Annual parental income in dollars

Figure 2 serves as a valuable reference for understanding the economic backgrounds of the students' families. It illustrates the various income levels of parents within the dataset, which can be essential for conducting analyses on the impact of parental income on high school students' college plans.

4. IQ: The IQ score of the student, determined through a recent test, with scores ranging from 60 to 140.

Figure 3 visually represents the distribution of IQ scores among high school students in the dataset. These IQ scores are derived from recent tests and indicate the student's cognitive abilities and intelligence. The chart illustrates the spread and concentration of these IQ scores, along with the defined range within which they fall.

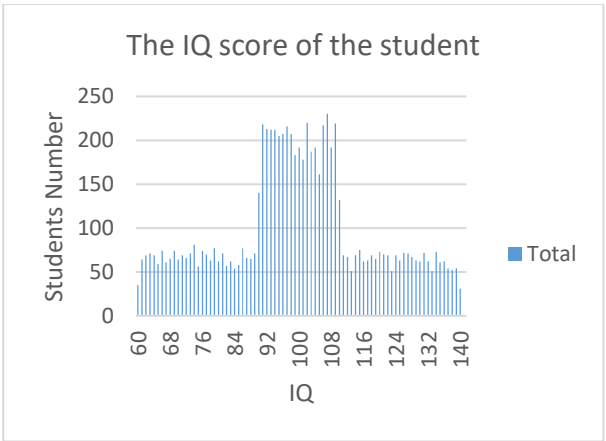


Figure 3: The IQ score of the student

Figure 3 offers valuable insights into the cognitive diversity of the student population, showcasing the distribution of IQ scores and highlighting the variability in intellectual abilities.

5. Encourage categorical variable indicating whether the parents encourage their child to pursue a college education, with two categories: "encourage" and "not encourage."

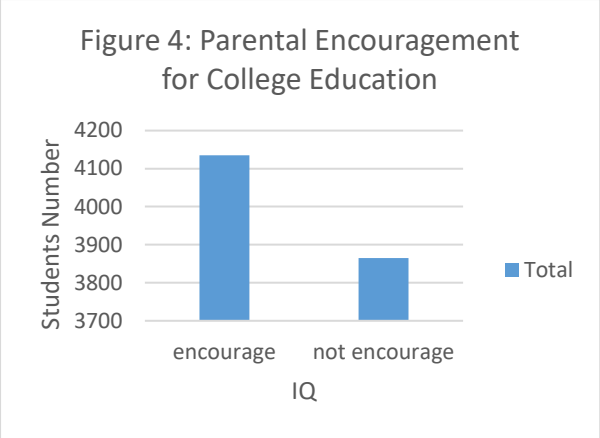


Figure 4: Parental Encouragement for College Education

In this categorical representation in Figure 4, "Encourage" denotes students whose parents actively support and encourage them to pursue a college education, totaling 4,135 students. Conversely, "Not Encourage" signifies students whose parents do not provide such active encouragement, with a count of 3,865 students. This figure offers a concise overview of the parental influence on students' college aspirations, highlighting the number of students falling into each category. It provides valuable insights into the role of parental encouragement in shaping educational decisions. It can inform strategies for supporting students with varying parental guidance in pursuing higher education.

Figure 5 shows the factors significantly influencing whether high school students plan to attend college. According to government officials, these four factors—Gender, Parent Income, IQ, and Encouragement—have been identified as the most influential in shaping students' college aspirations. The chart displays the count of students who fall into two distinct categories: those who do not plan to attend college and those who plan to attend college.

-Not plan (5404): This category represents the count of high school students who do not intend to attend college despite these influential factors. Specifically, there are 5,404 students in this category.

-Plan (2596): The adjacent category signifies the number of students planning to attend college despite the identified influential factors. The numeric value associated with this category is 2,596 students.

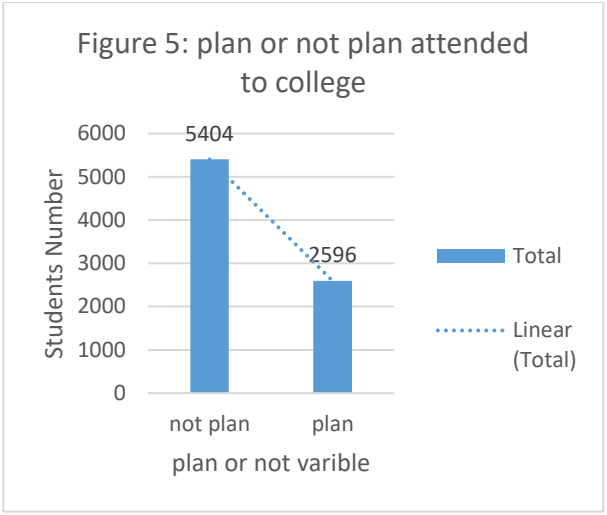


Figure 5: The IQ score of the student

Figure 5 offers a concise visual representation of the impact of these four influential factors on students' college plans. It reveals the distribution of students in terms of their intentions to attend college and provides valuable insights for educational planning and decision-making based on these influential variables.

3.2 Data Cleaning and Preparation

Data cleaning and preparation are essential stages in the data analysis, irrespective of the specific study or research field. These crucial steps serve several vital purposes. Firstly, they ensure data accuracy and reliability by identifying and rectifying errors, inconsistencies, and outliers that can significantly distort the analysis results. By addressing these issues, researchers can have confidence in the quality of the data they are working with.

Secondly, data cleaning and preparation enhance data consistency and uniformity. Datasets are often collected from diverse sources or multiple data collection points, resulting in variations in data formats and structures. Data can be more consistent through standardization and formatting, allowing for seamless analysis and comparisons [10].

Additionally, data cleaning and preparation help protect data privacy and confidentiality. Anonymization and desensitization are employed to safeguard sensitive information, ensuring that individuals' privacy rights are respected [11].

Furthermore, well-prepared data is more interpretable and user-friendly. Organizing and structuring data logically allows researchers to navigate and understand the dataset more easily, facilitating efficient analysis [12].

This study undertook several key steps to refine and structure the data effectively. First and foremost, measures were implemented to address potential data quality issues, which involved identifying and rectifying

missing data points and outliers that could significantly impact the accuracy of predictive models.
Furthermore, the dataset underwent a rigorous cleansing process to remove irrelevant or redundant information, streamlining it for analysis. For instance, students' names were desensitized and replaced with unique Student IDs to uphold privacy and data security standards, aligning with ethical considerations.
Feature selection and engineering were performed to enhance the dataset's utility for predictive modeling, which involved identifying the most relevant variables and creating new ones that could better capture the nuances of high school student's college plans.

Table 1 : High School Student Data Variables

Variable	Explanation	Range
Student ID	The unique identifying number of the student	1, 2, ..., 8000
Gender	The gender of the student	{male, female}
Parent income	The annual income of the parents, in US dollars	[4500, 82390]
IQ	The IQ of the student in the last test	[60, 140]
Encourage	Whether the parents encourage their child to go to college	{encourage, not encourage}
Plan	Whether the student eventually plans to go to college	{plan, not plan}

Table 1 is a reference for understanding the dataset's variables, meanings, and the permissible ranges or categories for each variable. These variables are crucial because government officials have identified them as the most influential factors when predicting whether high school students intend to enroll in college. Understanding these variables is essential for subsequent data analysis and predictive modeling to determine college plans among high school students.

4. METHODOLOGY

This section outlines the simplified methodology for predicting high school students' college plans utilizing a neural network.

4.1 Data Acquisition

The initial phase commences with acquiring a dataset encompassing extensive information about high school students. This dataset encompasses diverse attributes, including gender, IQ, parental income, and parental encouragement, in addition to the pivotal target variable, "College Plans." For this study, the dataset employed is sourced and supported by Kaggle, a renowned data-driven research and analysis platform.

4.2 Data Preprocessing

The data preprocessing procedure comprises a series of operations aimed at adapting the dataset to be suitable for training and assessing the performance of a neural network model. Let's delve into the key steps of this process:
In numerous machine learning and neural network models, categorical variables, including gender, parental encouragement, and college plans, need to be converted into a numerical representation. This conversion is crucial because the majority of algorithms operate with numerical data. A widely used technique known as one-hot encoding is applied to accomplish this. One-hot encoding generates binary columns for each distinct category within a categorical variable. A '1' in a particular binary column signifies the presence of that category, while a '0' denotes its absence. For example, consider "Gender" in Table 2, where categories would be transformed into two binary columns, simplifying further analysis.

Table 2 : Encoding of Categorical Variables

Gender		**encouragement**		**plan**	
Male	Female	encouragement	Not encouragement	plan	Not plan
1	0	1	0	1	0

Table 2 illustrates the one-hot encoding transformation applied to categorical variables, such as "Gender," "Encouragement," and "College Plans." The process creates binary columns, making the dataset more amenable to neural network analysis. Each binary column represents the presence or absence of a specific category within the original categorical variable, simplifying subsequent data processing and model training.

4.3 Feature Standardization

Standardizing features is a crucial preprocessing step that aims to rescale the values of different features in a dataset to a common scale. Specifically, it transforms the features to have a mean (average) value of 0 and a standard deviation (a measure of how spread out the values are) of 1. This process is essential, especially when working with neural networks, for several reasons:

- Consistent Scale: Standardization ensures that all features have the same scale, preventing certain features from dominating others during training. Without standardization, features with larger numerical values might disproportionately impact the model's learning.

- Optimized Training: Neural networks use optimization algorithms like gradient descent to adjust their parameters during training. Standardized features with a mean of 0 and a standard deviation of 1 make the optimization landscape more symmetric and well-behaved, leading to faster convergence and more stable training.
- Avoiding Vanishing or Exploding Gradients: In deep neural networks, gradients can become too small (vanishing gradients) or too large (exploding gradients) during backpropagation, making training difficult. Standardization helps mitigate these issues by keeping gradients within a reasonable range.

Standardization ensures that the features have a consistent scale and distribution, making it easier for neural networks to learn the underlying patterns in the data efficiently and effectively. Table 3 overviews the standard deviation (STD) for three key variables: Gender, Encouragement, and Plan. The standard deviation measures the amount of variation or dispersion in a dataset.

Table 3 : Standard Deviation of Gender, Encouragement, and Plan

	STD gender	STD encouragement	STD plan
male	0.96892	1.034275	0.693055
female	-1.031948	-0.96674	-1.44271

Table 3 illustrates the variability in gender, parental encouragement, and college plans among high school students, providing valuable insights into the dataset's characteristics and potential factors influencing college intentions. Positive and negative values indicate different levels of dispersion for each category within the variables

4.4 Feature Standardization

This study employed a Feedforward Neural Network (FNN) as our primary model for predicting high school students' college plans. FNNs are an ideal starting point for many classification problems, offering versatility and effectiveness [13]. These networks comprise an input layer, one or more hidden layers, and an output layer. The architecture's flexibility allows for experimentation with various configurations, including the number of neurons within each layer [14]. We proceeded to the model training phase once the neural network architecture was defined. During this stage, the neural network was exposed to the designated training dataset, where it underwent a learning process to discern intricate patterns and associations between the input features—such as gender, parental income, IQ, and parental encouragement—and the target variable of interest, "College Plans." This training process aimed to enable the model to make accurate predictions regarding high school students' intentions to pursue college education based on the provided input features.

4.5 Model Training

The neural network training phase is a critical step in the predictive modeling process. It involves exposing the model to the designated training dataset, which encompasses a comprehensive array of high school student data, including features like gender, parental income, IQ, and parental encouragement, as well as the crucial target variable, "College Plans." During this training phase, the neural network leverages its inherent capacity to discern intricate patterns and establish associations between these input features and the target variable, "College Plans." Through iterative adjustments and optimization of its internal parameters, the neural network strives to enhance its predictive capabilities, ultimately aiming to provide accurate predictions regarding high school students' intentions to pursue college education. This process is fundamental in enabling the model to generalize its learned patterns to new, unseen data, thereby facilitating its ability to predict college plans effectively.

In line with established best practices for model evaluation and validation, this study adopts a standard approach to data division [15-18]. The dataset is divided into two distinct subsets: a training set and a testing set. 70% of the data is allocated to the training set, while the remaining 30% is dedicated to the testing set. This data division serves a crucial purpose in the model development process. The training set trains the neural network, enabling it to learn patterns and associations within the data. Subsequently, the testing set, representing new, unseen data, is employed to rigorously assess the model's performance and ability to make accurate predictions regarding high school students' college plans. This clear differentiation between training and testing sets ensures the model's predictive capabilities are rigorously and objectively evaluated, providing valuable insights into its generalization performance and overall effectiveness in real-world scenarios.

4.5 Run the model

To execute the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model using Python Algorithm1, the following Python code presents a concise overview of the code's functionality. It showcases the implementation of a Feedforward Neural Network (FNN) for predicting high school students' college plans based on a CSV dataset. The code leverages sci-kit-lear for machine learning and

pandas for data handling. Before running the code, ensure that you have installed these libraries.

Algorithm1: AIRPCP using Feedforward Neural Network (FNN) Classification

```
1. import pandas as pd
2. from sklearn.model_selection import
   train_test_split
3. from sklearn.preprocessing import
   StandardScaler
4. from sklearn.neural_network import
   MLPClassifier
5. from sklearn.metrics import accuracy_score,
   classification_report
6. # Load the dataset from a CSV file
7. data = pd.read_csv('your_dataset.csv')
8. # Define features (X) and target variable (y)
9. X = data[['Gender', 'Parent_income', 'IQ',
   'Encourage']]
10. y = data['College_Plans']
11. # Perform one-hot encoding for categorical
    variables
12. X = pd.get_dummies(X, columns=['Gender',
    'Encourage'], drop_first=True)
13. # Split the data into training and testing sets
    (70% training, 30% testing)
14. X_train, X_test, y_train, y_test =
    train_test_split(X, y, test_size=0.3,
    random_state=42)
15. # Standardize features to have mean=0 and std=1
16. scaler = StandardScaler()
17. X_train = scaler.fit_transform(X_train)
18. X_test = scaler.transform(X_test)
19. # Create and train the Feedforward Neural
    Network (FNN) model
20. model = MLPClassifier(hidden_layer_sizes=(50,
    50), max_iter=1000, random_state=42)
21. model.fit(X_train, y_train)
22. # Make predictions on the test data
23. y_pred = model.predict(X_test)
24. # Evaluate the model's performance
25. accuracy = accuracy_score(y_test, y_pred)
26. report = classification_report(y_test, y_pred)
27. # Print the accuracy and classification report
28. print(f'Accuracy: {accuracy:.2f}')
29. print('Classification Report:')
30. print(report)
```

- Explanation of the provided code:

1. Load the dataset from a CSV file.
2. Define the features (X) and the target variable (y).
3. Perform one-hot encoding for categorical variables ('Gender' and 'Encourage').
4. Split the data into training and testing sets (70% training and 30% testing).
5. Standardize the features with a mean of 0 and a standard deviation 1.
6. Create an FNN model with two hidden layers, each containing 50 neurons.
7. Train the FNN model on the training data.
8. Make predictions on the test data.
9. Evaluate the model's performance using accuracy and

5. RESULT

The predictive model results for high school students' college plans indicate promising performance. The accuracy is approximately 84.75%, signifying the proportion of correct predictions out of the total predictions made. The classification report delves deeper into the model's performance by examining precision, recall, and score for each class.

For students with no plans to attend college (Class 0), the model achieved a precision of 78%, implying that 78% of the predicted "no college plans" instances were correct. The recall for this class is 73%, indicating that the model correctly identified 73% of all the actual "no college plans" cases. The F1 score, which combines precision and recall into a single metric, stands at 0.75 for this class.

On the other hand, for students with intentions to attend college (Class 1), the model demonstrated a precision of 88%, indicating that 88% of the predicted "college plans" instances were correct. The recall for this class is 90%, signifying that the model correctly identified 90% of all the actual "college plans" cases. The F1-score for this class is notably higher at 0.89, reflecting the model's ability to classify students planning to attend college effectively.

In summary, the model exhibits strong predictive capabilities, particularly in identifying students with plans to attend college, where it achieves higher precision and recall. These results underscore the potential of this predictive model in assisting educational institutions and policymakers in addressing the educational aspirations of high school students.

Table 4 provides an overview of the performance metrics for the predictive model used to forecast high school student's college plans. The model's accuracy, precision, recall, and F1 score are reported for each class within the target variable.

Table 4 provides an overview of the performance metrics for the predictive model used to forecast high school student's college plans. The model's accuracy, precision, recall, and F1 score are reported for each class within the target variable.

Table 4 : AIRPCP Model Classification Report

	precision	recall	f1-score	support
0	0.78	0.73	0.75	766
1	0.88	0.9	0.89	1634
accuracy			0.85	2400
Macro avg.	0.83	0.82	0.82	2400

Weighted avg.	0.85	0.85	0.85	2400
Accuracy: 84.75%				

5.1 Hyperparameter Tuning

Hyperparameter tuning, handling imbalanced datasets, and conducting extensive data preprocessing and feature engineering are crucial to enhancing predictive accuracy and relevance in machine learning models [19]. While the provided code serves as a simplified illustration, you can incorporate these advanced techniques into the AIRPCP model by

To perform hyperparameter tuning, use libraries like scikit-learn's `GridSearchCV` or `RandomizedSearchCV` [20]. These tools allow you to systematically search for the best hyperparameters for neural networks, such as the learning rate, number of hidden layers, and neurons per layer.

Enhancing predictive accuracy and model relevance in machine learning [19] involves pivotal steps. Integrate advanced techniques into the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model to enable a systematic exploration of hyperparameter combinations encompassing learning rates, hidden layer configurations, and neuron counts per layer, leading to the discovery of optimal neural network settings. Utilize libraries in Python such as sci-kit learn GridSearchCV or RandomizedSearchCV to facilitate this exploration [20].

By changing the 'hidden_layer_sizes in Algorithm1 line 20 to be 'hidden_layer_sizes': [(50, 50), (100, 100), (50, 50, 50)], 'alpha': [0.0001, 0.001, 0.01], 'learning_rate_init': [0.001, 0.01, 0.1]. Table 4 presents the classification report after fine-tuning the neural network to optimize model performance by adjusting hidden layer sizes.

Table 5 : AIRPCP Model Classification Report

	precision	recall	f1-score	support
0	0.77	0.74	0.76	766
1	0.88	0.9	0.89	1634
accuracy			0.85	2400
macro avg	0.83	0.82	0.82	2400
weighted avg	0.85	0.85	0.85	2400
Accuracy: 84.88%				

Table 5 demonstrated improvements in AIRPCP Model performance. In this updated model, we achieved an accuracy of 84.88%, a slight increase compared to the initial model's accuracy of 84.75%. These results highlight the effectiveness of hyperparameter tuning in enhancing the model's predictive capabilities.

In terms of precision, there are consistently high values for both classes. Class 0 (No College Plans) maintains a precision of 77%, indicating that 77% of the predictions for this class were correct. In comparison, Class 1 (College Plans) exhibits a precision of 88%, demonstrating the model's accuracy in identifying students with intentions to attend college.

Additionally, the recall values remain strong. Class 0 boasts a recall of 74%, indicating that the model correctly identified 74% of all instances where students had no college plans. Class 1 exhibits an even higher recall of 90%, reflecting the model's ability to capture students with college plans effectively.

The F1 scores for both classes also improved, with Class 0 achieving a score of 0.76 and Class 1 reaching an impressive F1 score of 0.89. These scores represent a harmonious balance between precision and recall, signifying the model's ability to make accurate predictions while minimizing false positives and false negatives.

The hyperparameter-tuned model has demonstrated superior performance, resulting in slightly higher accuracy and refined precision-recall trade-offs. These enhancements signify the value of optimizing neural network hyperparameters for better predictive accuracy and relevance in forecasting high school student's college plans.

5.2 Visualizations and charts illustrating the result

- Model performance metrics

Visualizations and charts are pivotal in conveying the outcomes and insights derived from the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model's classification report, especially after fine-tuning its hyperparameters. The visual representation in Figure 6 provides a brief and understandable means of interpreting complex performance metrics.

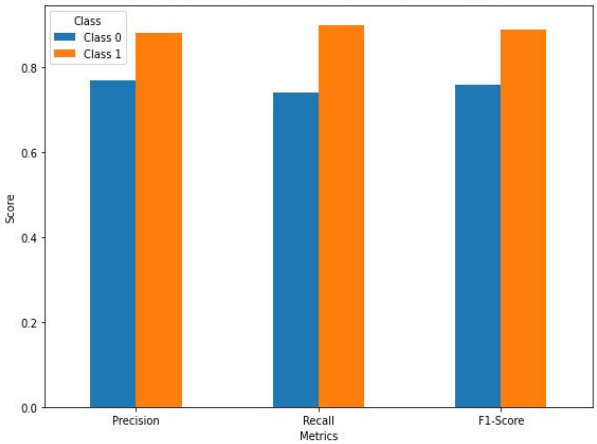


Figure 6: AIRPCP Model Performance Metrics by Class

Figure 6 employs visual aids to present the model's precision, recall, and F1 score for each class (Class 0 and Class 1) in a bar chart format. This graphical representation allows for a clear comparison of the model's performance metrics across different classes, providing stakeholders with an intuitive understanding of how the model discriminates between the two target categories. These charts are valuable tools for decision-makers, educators, and researchers to gain insights into the model's ability to accurately predict high school students' college plans and make informed decisions based on its performance.

- Confusion matrix

A confusion matrix is a table or matrix used in machine learning and classification to evaluate the performance of a classification model, particularly in binary classification problems (problems with two classes or categories). It is a crucial tool for understanding how well a model makes correct and incorrect predictions.

A confusion matrix provides a comprehensive summary of a classification model's performance, allowing practitioners to understand where it makes correct predictions and where it may need improvement. Table 5 shows the confusion matrixIn this table, two rows labeled "Actual Positive" and "Actual Negative" correspond to the true class labels of the instances in the dataset. The columns "Predicted Positive" and "Predicted Negative" represent the model's predictions for each instance. The numbers within each table cell reflect the count of instances that belong to specific combinations of actual and predicted classes. For instance, the top-left cell shows the count of truly positive instances correctly predicted as positive. In contrast, the bottom-right cell indicates the count of truly negative instances and correctly predicted as negative.

Table 5 : Confusion matrix

	Actual Positive	Actual Negative
Predicted Positive	569	166
Predicted Negative	197	1468

This visual representation aids in assessing the model's accuracy, precision, recall, and overall performance in making binary classification decisions.

Also, Figure 7 displays the confusion matrix as visualization; the figure serves as a concise summary of how well the model has made predictions in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The upper-left quadrant represents TP, indicating instances where the model correctly identified positive cases. In contrast, the upper-right quadrant represents FP, instances where the model incorrectly predicted positive cases. The lower-left quadrant symbolizes FN, indicating cases where the model failed to identify positive instances. Finally, the lower-right quadrant signifies TN, showcasing instances where the model correctly recognized negative cases. By examining the values in each cell of this matrix, one can gain insights into the model's strengths and weaknesses.

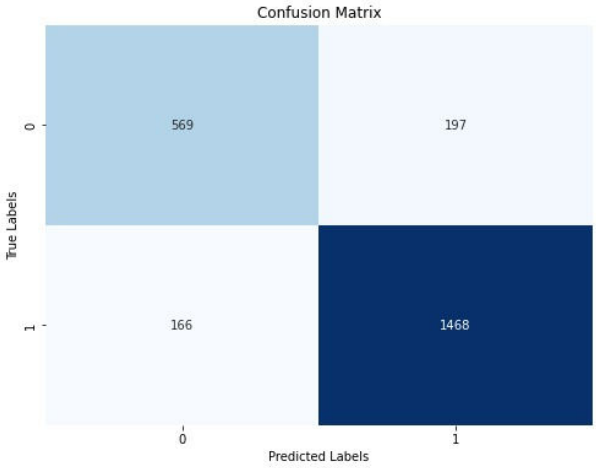


Figure 7: confusion matrix as visualization

1. True Positives (TP): The top-left cell (569) represents the number of instances where the model correctly predicted that high school students have college plans (the positive class).
2. False Negatives (FN): The top-right cell (197) represents the number of instances where the model incorrectly predicted that high school students do not have college plans when they do. In other words, it's the number of students the model missed in predicting as having college plans.
3. False Positives (FP): The bottom-left cell (166) represents the number of instances where the model incorrectly predicted that high school students have college plans when they do not. It's the number of students falsely classified as having college plans.
4. True Negatives (TN): The bottom-right cell (1468) represents the number of instances where the model correctly predicted that high school students do not have college plans (the negative class).
5. These values provide insight into the model's performance, particularly its ability to distinguish between students with and without college plans

6. DISCUSSION

The results of this study reveal the potential and effectiveness of the AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model in predicting high school students' college plans. The model demonstrates promising performance, achieving an accuracy rate of 84.88%, which signifies the proportion of correct predictions out of the total predictions made.

A deeper analysis of the model's performance is provided through the classification report. The report dissects the model's precision, recall, and F1 score for each class, offering valuable insights into its predictive capabilities. For students with no plans to attend college (Class 0), the model achieved a precision of 78%, indicating that 78% of the predicted "no college plans" instances were correct. The recall for this class is 73%, signifying that the model correctly identified 73% of all the actual "no college plans" cases. The F1 score for this class stands at 0.75, combining precision and recall into a single metric.

In contrast, for students with intentions to attend college (Class 1), the model demonstrates a precision of 88%, indicating that 88% of the predicted "college plans" instances were correct. The recall for this class is even higher at 90%, signifying that the model correctly identified 90% of all the actual "college plans" cases. The F1-score for this class impressively reaches 0.89, reflecting the model's ability to classify students planning to attend college effectively.

Overall, the model exhibits strong predictive capabilities, particularly in identifying students with plans to attend college, where it achieves higher precision and recall. These results underscore the potential of this predictive model in assisting educational institutions and policymakers in addressing the educational aspirations of high school students.

The presentation of these results is further enriched by Table 3, which provides an overview of the performance metrics for the predictive model, including accuracy, precision, recall, and F1-score, for each class within the target variable. This tabular format summarizes the model's performance, facilitating easy comparison and evaluation of its predictive power.

Hyperparameter tuning, an essential component of model optimization, is highlighted in section 5.1. The study emphasizes the significance of this step in enhancing predictive accuracy and relevance. The model performs refined by systematically exploring hyperparameter combinations, including learning rates, hidden layer configurations, and neuron counts per layer. Utilizing libraries such as scikit-learn's GridSearchCV or RandomizedSearchCV enables a comprehensive search for the best hyperparameters. This process improves accuracy from 84.75% to 84.88%, demonstrating the tangible benefits of hyperparameter tuning.

The improvement in precision and recall is consistent for both classes. Class 0 (No College Plans) maintains a precision of 77%, indicating the model's accuracy in identifying students without college plans. Class 1 (College Plans) exhibits a precision of 88%, highlighting the model's precision in recognizing students with intentions to attend college. The recall values remain strong, with Class 0 at 74% and Class 1 at an impressive 90%. These enhancements are further reflected in the F1 scores, which achieve a harmonious balance between precision and recall.

In summary, the hyperparameter-tuned model presents superior performance with a slightly higher accuracy and refined precision-recall trade-offs. These improvements underscore the value of optimizing neural network hyperparameters to achieve enhanced predictive accuracy and relevance in forecasting high school student's college plans.

The discussion extends to the presentation of visualizations and charts in section 5.2, which are crucial in conveying model outcomes. As seen in Figure 6, visual representation offers an intuitive means of interpreting complex performance metrics. The bar chart format of Figure 6 presents precision, recall, and F1 score metrics for each class, facilitating a clear comparison of the model's performance across different categories. These visual aids empower stakeholders, including decision-makers, educators, and researchers, to gain insights into the model's ability to accurately predict high school students' college plans. The confusion matrix, both in table form and as a visualization (Figure 7), is a pivotal tool for evaluating the model's performance, particularly in binary classification problems. This comprehensive summary allows practitioners to assess where the model excels and where improvements are needed. One can understand the model's strengths and weaknesses by examining true positives, false negatives, false positives, and true negatives.

6. CONCLUSION AND FUTURE WORK

The AIRPCP (Artificial Intelligence for Educational Planning of College Pursuits) model significantly advances educational data analytics. This research has demonstrated the model's ability to accurately predict high school students' college plans, offering valuable insights into their educational aspirations.

The AIRPCP model can be a valuable tool for educators, policymakers, and researchers to identify students at risk of not attending college or dropping out of high school, enabling the development of targeted interventions to support their educational goals. Additionally, it can be a valuable tool for evaluating the effectiveness of college readiness programs and other educational initiatives, providing data-driven insights to inform educational policy and practice.

One crucial area of future research is addressing data limitations. Expanding the dataset to encompass a more diverse and representative sample of high school students from various demographics and regions can enhance the model's generalizability. Additionally, collecting data on other relevant factors, such as socioeconomic background, extracurricular activities, and geographic location, can contribute to a more holistic understanding of students' college plans.

Another promising direction for future research is exploring the temporal dynamics of students' educational aspirations. Longitudinal data tracking students' plans and their evolution over time can provide insights into the changing nature of college aspirations and the factors influencing these changes.

Finally, collaboration with educational institutions and policymakers is vital in translating research findings into actionable strategies. Future research can involve partnerships with schools and education departments to implement and evaluate the AIRPCP model in real-world settings. This practical application can help refine the model further and tailor it to the specific needs of educators and students.

The AIRPCP model can be further developed into a proactive tool for educational guidance and support by addressing these areas of future research. It enables educators to identify and assist students at risk of falling behind on their educational journey. Ultimately, future research in this domain holds exciting possibilities for advancing the accuracy and applicability of predictive models like AIRPCP to promote college access and attainment for all students.

REFERENCES

[1] Rozental, A., Forsström, D., Hussoon, A., & Klingsieck, K. B. (2022). Procrastination among university students: differentiating severe cases in need of support from less severe cases. Frontiers in psychology, 13, 783570.

[2] Adeyemo, D. A., & Jegede, D. J. (2023). Sociopsychological determinants of career maturity among secondary school students in Osogbo, Osun State, Nigeria. Journal of Psychological Perspective, 5(1), 9-16.

[3] Kuroki, M. (2023). Integrating data science into an econometrics course with a Kaggle competition. The Journal of Economic Education, 1-15.

[4]. Chang, L., Wang, Y., Liu, J., Feng, Y., & Zhang, X. (2023). Study on factors influencing college students' digital academic reading behavior. Frontiers in psychology, 13, 1007247.

[5] Pan, X. S., Li, C., & Watts, T. W. (2023). Associations between preschool cognitive and behavioral skills and college enrollment: Evidence from the Chicago School Readiness Project. Developmental Psychology, 59(3), 474.

[6] Ye, X. (2023). Improving College Choice in Centralized Admissions: Experimental Evidence on the Importance of Precise Predictions. Education Finance and Policy, 1-75.

[7] Yan, J., & Chi, X. (2023, April). Analysis and Prediction of College Students' Employment based on Decision Tree Classification Algorithm. In 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-6). IEEE.

[8] Jiao, P., Ouyang, F., Zhang, Q., & Alavi, A. H. (2022). Artificial intelligence-enabled prediction model of student academic performance in online engineering education. Artificial Intelligence Review, 55(8), 6321-6344.

[9] Gumba, G., & Paragas, J. R. (2022, September). Prediction Analysis Of Student Admission To Information Technology Education (ITE) Programs Using Classification Algorithm. In 2022 Fifth International Conference on Vocational Education and Electrical Engineering (ICVEE) (pp. 112-117). IEEE.

[10] Singh, G., Singh, J., & Prabha, C. (2022, June). Data visualization and its key fundamentals: A comprehensive survey. In 2022 7th international conference on communication and electronics systems (ICCES) (pp. 1710-1714). IEEE.

[11] Murugeshwari, B., Selvaraj, D., Sudharson, K., & Radhika, S. (2023). Data Mining with Privacy Protection Using Precise Elliptical Curve Cryptography. Intelligent Automation & Soft Computing, 35(1).

[12] Bharadiya, J. P. (2023). Leveraging Machine Learning for Enhanced Business Intelligence. INTERNATIONAL JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY, 7(1), 1-19.

[13] Ren, Y. M., Alhajeri, M. S., Luo, J., Chen, S., Abdullah, F., Wu, Z., & Christofides, P. D. (2022). A tutorial review of neural network modeling approaches for model predictive control. Computers & Chemical Engineering, 107956.

[14] Siłka, J., Wieczorek, M., & Woźniak, M. (2022). Recurrent neural network model for high-speed train vibration prediction from time series. Neural Computing and Applications, 34(16), 13305-13318.

[15] Ahmed, N., Hoque, M. A. A., Arabameri, A., Pal, S. C., Chakrabortty, R., & Jui, J. (2022). Flood susceptibility mapping in Brahmaputra floodplain of Bangladesh using deep boost, deep learning neural network, and artificial neural network. Geocarto International, 37(25), 8770-8791.

[16] Hakim, W. L., Nur, A. S., Rezaie, F., Panahi, M., Lee, C. W., & Lee, S. (2022). Convolutional neural network and long short-term memory algorithms for groundwater potential mapping in Anseong, South Korea. Journal of Hydrology: Regional Studies, 39, 100990.

[17] Guo, Y., Yang, D., Zhang, Y., Wang, L., & Wang, K. (2022). Online estimation of SOH for lithium-ion battery based on SSA-Elman neural network. Protection and Control of Modern Power Systems, 7(1), 40.

[18] Samhan, L. F., Alfarra, A. H., & Abu-Naser, S. S. (2022). Classification of Alzheimer's disease using convolutional neural networks.

[19] Hakak, S., Alazab, M., Khan, S., Gadekallu, T. R., Maddikunta, P. K. R., & Khan, W. Z. (2021). An ensemble machine learning approach through effective feature extraction to classify fake news. Future Generation Computer Systems, 117, 47-58.

[20] Wade, C., & Glynn, K. (2020). Hands-On Gradient Boosting with XGBoost and scikit-learn: Perform accessible machine learning and extreme gradient boosting with Python. Packt Publishing Ltd product quality dimensions on improving the order-winners and customer satisfaction," *Int. J. Product. Qual. Manag.*, vol. 36, no. 2, pp. 169–186, 2022, doi: 10.1504/IJPQM.2021.10037887.

Journal of Intelligent System and Applied Data Science (JISADS)

Journal homepage : https://www.jisads.com

ISSN (2974-9840) Online

ENHANCING UNIVERSITY ELECTRONIC BOOK ACQUISITION STRATEGY USING A DEEP FOREST FUSION APPROACH

Sanjai Kumar. T[1] *Mohammad Sameer Aloun*[2*]

[1]*Periyar Maniammai Institute of Science and Technology, Thanjavur, Tamil Nadu, India*

[2]*Faculty of Science and Information Technology Irbid National University*

Abstract

This study presents a new approach, LHGCAT-XDF, to improve the efficiency of electronic book procurement in university settings by combining the strengths of the LightGBM and CatBoost algorithms. This innovative model benefits from the LightGBM's minimal memory usage and CatBoost's reduced time complexity. Through testing, it's shown that LHGCAT-XDF surpasses standard machine learning models in overall effectiveness, successfully addressing the shortcomings of conventional procurement strategies in terms of accuracy and efficiency. Thus, it offers dependable guidance for the selection of electronic books in university libraries.

Keywords : Machine learning, LightGBM, CRTF, CatBoost

1. INTRUDUCTION

In the rapidly evolving information society and with the widespread adoption of mobile devices, the borrowing volume of physical books has been on a consistent decline. Concurrently, the demand for electronic books among readers has been increasing. This trend not only fosters a growing demand for a variety of electronic books but also sets higher standards for their quality and the services provided.

The primary goal of constructing university libraries is to meet readers' demands for more accessible and diverse academic resources, thereby enhancing service quality to align with the trends of the information age. This entails not just updates to physical infrastructure and spatial

optimization but also improvements in the volume and quality of literature resources and an overall enhancement of library services [1]. In this transformation, libraries must evolve from traditional service models to smarter, more reader-centric approaches. However, most domestic university libraries still adhere to conventional book procurement strategies, relying on annual budgets, the experience of purchasers, recommendations from faculty and students, and suggestions from vendors to compile their procurement lists [2].

While some university libraries have begun to employ information technology to develop decision-support systems for book procurement, most systems still base their purchasing decisions on existing collection catalogs, borrowing data, and reader information, using statistical analysis [3]. Given the dynamic and uncertain nature of this data, understanding readers' reading needs and purchasing books that best meet these needs within a limited budget remains a key challenge in the book procurement process.

To enhance the efficiency and quality of electronic book procurement in university libraries and to explore the complex and variable relationship between book attributes and reader demands, this article adopts a hybrid deep forest model for predicting electronic book procurement in university libraries. This model not only significantly improves prediction accuracy compared to traditional machine learning models but also reduces the time complexity of model predictions and the difficulty of tuning hyperparameters, making it a more accurate and efficient algorithm for the field of book procurement prediction.

2. DEEP FOREST MODEL

The Deep Forest model, introduced by Zhou Zhi-Hua and Feng Jie in 2019, represents an ensemble method based on decision trees, falling under the umbrella of decision tree ensemble techniques [4]. Unlike Deep Neural Networks (DNNs), Deep Forest showcases superior competitiveness by requiring fewer hyperparameter adjustments, thereby reducing the time cost associated with hyperparameter tuning. It adapts well to datasets of various sizes and exhibits excellent generalization capabilities. These advantages have led to its wide application across different fields, affirming its robustness in classification and prediction tasks.

The Deep Forest model consists of two main components: Multi-Grained Scanning and Cascade Forest. Multi-Grained Scanning aims to analyze input features to unearth the sequential relationships between them. This process involves scanning the input feature vector with sliding windows of various lengths, generating multiple k-dimensional feature fragments [5]. These fragments are then fed into Random Forest (RF) and Completely Random Tree Forest (CRTF) models, with their class probability vectors, concatenated to form a transformed feature vector for the Cascade Forest input.

Cascade Forest, structured in multiple levels, each contains several ensemble learning classifiers, such as decision tree forests, XGBoost, LightGBM, or CatBoost. This hierarchical organization aims to build a stronger ensemble with better generalization performance. The model's design allows flexible feature learning and combination, with k-fold cross-validation employed in training each forest to prevent overfitting. The cascade structure dynamically adjusts the number of levels based on the training process, enhancing model complexity adaptability and training loss control.

To further enhance the performance of Deep Forest on smaller datasets, this paper introduces optimizations through LightGBM and CatBoost algorithms, simplifying the Multi-Grained Scanning structure and optimizing the number of random forests [6],. LightGBM reduces the number of data instances with minor gradients by utilizing one-sided gradient sampling, thus saving time and space [7]. CatBoost, through an optimized gradient boosting method and combining symmetric tree models with feature quantile metrics, simplifies model training and reduces data preprocessing complexity, offering an efficient and precise solution for electronic book procurement prediction, particularly with text-type electronic book attributes [8].

3. DEVELOP AN E-BOOK FORECASTING MODEL USING AN OPTIMIZED DEEP FOREST ALGORITHM.

Utilizing the past five years of access records from S Academy Library as a foundation, this study constructs a precise model for predicting e-book procurement. Initial steps involve preprocessing the gathered data and employing value indicators to sift through e-book interview decision influencers. The BM25 algorithm [9] facilitates the engineering of text features, with the refined samples then

applied to develop the prediction model using the LHGCAT-XDF approach, as illustrated in Figure 1.

Feature Analysis

Diverse decision-making elements influence e-book interviewers across university libraries, with varying degrees of importance attached to each. An exhaustive analysis, incorporating practical insights from the e-book acquisition processes in numerous university libraries and academic perspectives, leads to a detailed summation of the current influential factors in library acquisition decisions. Following this, characteristic variables are identified and quantified via information gain to aid in constructing the subsequent e-book prediction model [10].

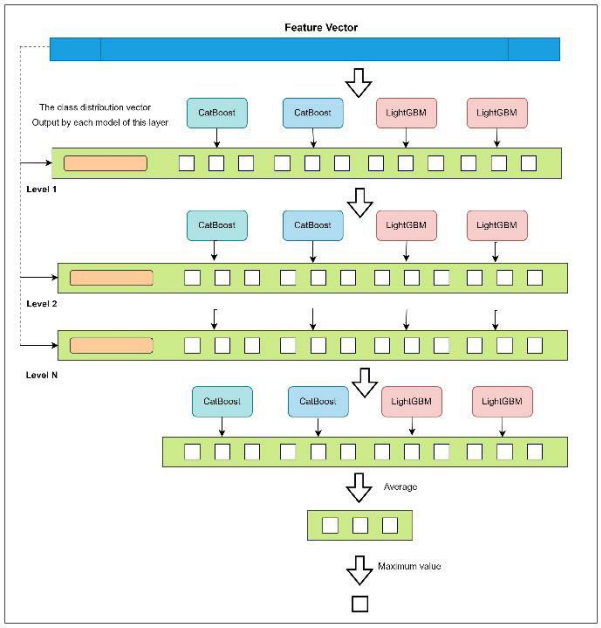


Fig. 1: LHGCAT-XDF algorithm diagram

Influencing Factors

The data derived from reader interactions with e-books on platforms like library portals encompass both basic (personal and borrowing history) and behavioral (search activities and database access) information. This amalgamated data aids in sculpting a comprehensive reader profile, essential for informed e-book purchasing decisions, thereby addressing both explicit and latent reader needs [11] as shown in table 1.

Data Preprocessing

Preprocessing entails organizing the collected data—ranging from e-book details and reader feedback to operational logs and financial records—filtering out pertinent information for dataset creation. The approach involves various technical methods, including web crawlers, to fill in missing values and employs BM25 for correlating words with documents, thus laying the groundwork for the model [12].

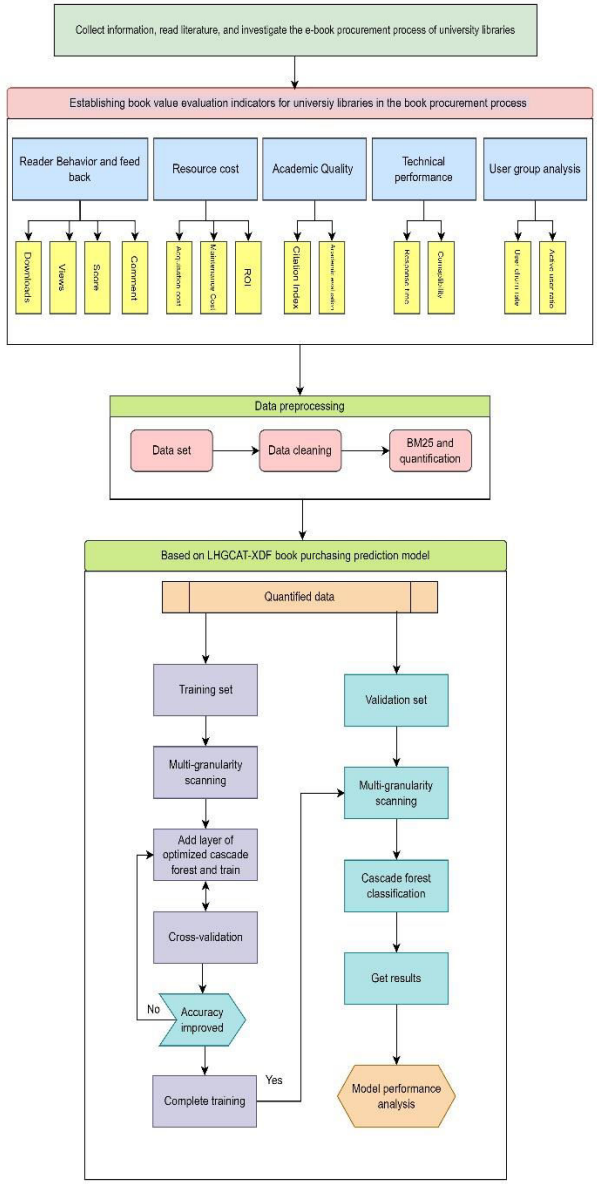


Figure. 2: Flow chart of e-book purchasing prediction model

Model Prediction and Evaluation

The LHGCAT-XDF-based e-book purchase prediction model operates by initially splitting cleaned data into training and validation sets. Feature selection through multi-granularity scanning precedes training with LightGBM and CatBoost models, incorporating 10-fold cross-validation. The procedure iterates until no significant accuracy improvements are observed. Model performance is assessed using metrics like Accuracy, Precision, Recall, Specificity, and F1 score [13].

Assessment Metrics

When assessing model performance, metrics such as Accuracy, Precision, Recall, Specificity, and F1 score (refer to Table 2 for details) are typically utilized to comprehensively gauge the model's effectiveness

Table 1. The influencing factors in electronic book procurement

Influencing factors	Metrics	Data Sources	Method of Obtaining
user behavior	Downloads Views score Comment	E-book platform backend statistics E-book platform backend statistics e-book platform User review feedback	Download statistics provided by e-book platforms Number of user views recorded through the e-book platform Get ratings through the platform Get feedback from user reviews
resource cost	acquisition cost Maintenance cost ROI	Actual cost of purchasing/subscribing to Maintenance costs of e-book library operations ROI	Obtained from financial records or purchase contracts Obtained from financial records or operating expense details Calculate the benefit and cost ratio of e-book access to obtain
academic quality	citation index academic evaluation	Data from academic databases or citation tools Academic review results	Access through academic databases, citation tools, etc. Obtain from relevant academic publications, journals or platforms
Technical performance	Response time compatibility	E-book platform performance monitoring Platform test report, user feedback	Obtained through performance monitoring tools By conducting platform testing and collecting user feedback
User group analysis	User churn rate Active user ratio	User behavior data analysis tool E-book platform backend statistics	Calculated through user behavior data analysis tools By counting the ratio of the number of active users to the total number of users

Among these, TP (True Positive) represents correctly identified positive instances, TN (True Negative) denotes accurately identified negative instances, FP (False Positive) signifies incorrectly identified positive instances, and FN (False Negative) indicates erroneously identified negative instances [14].

Table 2. Model evaluation criteria

Evaluation Criteria	Meaning	Formula
Accuracy	Represents the ratio of the number of samples that predict the entire sample correctly to the number of the population	TP + TN / TP + TN + FP + FN
Precision	Indicates the proportion of classified positive samples to all samples classified as positive	TP / TP + FP
Recall	Indicates the probability of predicting a positive	TP / TP + FN

	sample among all positive samples	
Specificity	Indicates the proportion of samples correctly predicted as negative class to all actual negative class samples.	TN / TN + FP
F1 value	Expressed as the weighted average of precision and recall	2 x Precision x Recall / Precision + Recall

Experimental Design and Result Analysis

The construction of the deep forest model constitutes a crucial phase, with particular emphasis placed on forest establishment. To enhance model accuracy, various forest parameters require iterative adjustments. The paper introduces LightGBM and XGBoost into the cascade forest structure. LightGBM offers diverse parameter configurations for optimization via cross-validation, where Learning_rate denotes the model's learning rate. A higher learning rate facilitates faster descent along the loss gradient, and vice versa. Num_leaves signifies the number of leaves on each tree, and Max_depth sets the maximum depth of the decision tree regression model. Feature_fraction subsamples features to expedite training and prevent overfitting. Refer to Table 3 for parameter specifications.

Table 3. LightGBM classification model parameter settings

Parameter	Numerical Value	Parameter	Numerical Value
Learning_rate	0.005	Feature_fraction	0.8
N_estimator	927	Num_leaves	10
Max_depth	-1	Max_bin	245
Bagging_fraction	0.6	Bagging_freq	0

By fine-tuning these parameters, the cascade forest was restructured. Multiple experiments were conducted to strike a balance between model runtime and accuracy. Ultimately, the parameter N_estimators = 927 was chosen, resulting in a model accuracy of 79%.

Model Comparison

To underscore the predictive superiority of the deep forest model, traditional learning models (LightGBM, Random Forest, KNN, CNN) were employed to predict sample data, and each model's assessment metrics were compared (refer to Table 4). In terms of specific values, the deep forest achieves an accuracy of 79.0%, markedly surpassing other models. Furthermore, Deep Forest's accuracy of 83.72% also outperforms other models. Recall, specificity, and F1 score also exhibit a notably favorable trend for the deep forest. Although traditional machine learning models boast shorter runtimes, their various assessment metrics pale in comparison to the results obtained with deep forests.

Table 4. Performance evaluation table of various models

Model	LightGBM	random forest	KNN	CNN	LHGCAT-XDF
Accuracy	71.09%	71.5%	69.5%	72.5%	79.70%
Precision	77.63%	72.26 %	76.71 %	77.78 %	83.72%
Recall	0.59	0.61	0.65	0.63	0.90
Specificity	0.83	0.82	0.83	0.82	0.86
F1-Score	67.05%	68.16 %	64.74 %	69.61 %	77.42%

4. CONCLUSION

Accurately predicting electronic book procurement holds paramount importance for university library development. However, existing prediction models suffer from issues of simplicity and low accuracy. To address this, we propose the LIGHT-XDF algorithm, a deep forest algorithm leveraging LightGBM and CatBoost. LightGBM and CatBoost are integrated into the cascade forest, with CatBoost enhancing prediction accuracy and LightGBM reducing model complexity. The LIGHT-XDF algorithm utilizes reader behavioral data and electronic library collection data for purchase predictions. Experimental findings demonstrate that, compared to alternative models, LIGHT-XDF exhibits superior overall performance. Future endeavors will focus on validating the robustness and generalization capability of the LIGHT-XDF algorithm through extensive performance testing on diverse library

collection datasets. Additionally, exploration of various new technologies will be pursued to enhance the accuracy of overall e-book procurement predictions.

References

[1] Li, M. (2007). Application of web-based data mining technology in digital libraries. Journal of Academic Library and Information Science, 25, 44-46.

[2] Affum, M. Q. (2023). Book Acquisition in the Modern University Library: Challenges and Opportunities. Library Philosophy & Practice.

[3] Anna, N. E. V., & Mannan, E. F. (2020). Big data adoption in academic libraries: a literature review. Library Hi Tech News, 37(4), 1-5.

[4] Zhou, Z. H., & Feng, J. (2019). Deep forest. National science review, 6(1), 74-86.

[5] Ma, W., Yang, H., Wu, Y., Xiong, Y., Hu, T., Jiao, L., & Hou, B. (2019). Change detection based on multi-grained cascade forest and multi-scale fusion for SAR images. Remote Sensing, 11(2), 142.

[6] Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisciplinary review. Journal of big data, 7(1), 94.

[7] Liang, Y., Wu, J., Wang, W., Cao, Y., Zhong, B., Chen, Z., & Li, Z. (2019, August). Product marketing prediction based on XGboost and LightGBM algorithm. In Proceedings of the 2nd international conference on artificial intelligence and pattern recognition (pp. 150-153).

[8] Jabeur, S. B., Gharib, C., Mefteh-Wali, S., & Arfi, W. B. (2021). CatBoost model and artificial intelligence techniques for corporate failure prediction. Technological Forecasting and Social Change, 166, 120658.

[9] Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. Foundations and Trends® in Information Retrieval, 3(4), 333-389.

[10] Zhao, F., Kumamoto, E., & Yin, C. (2021, July). The effect and contribution of e-book logs to model creation for predicting students' academic performance. In 2021 International Conference on Advanced Learning Technologies (ICALT) (pp. 187-189). IEEE.

[11] Trakarnsakdikul, N., Chaiyaphan, S., & Leecharoen, B. (2021). Factors Affecting E-Book Purchase Decisions of Customers in Thailand. Asian Administration & Management Review, 4(1).

[12] Whissell, J. S., & Clarke, C. L. (2011). Improving document clustering using Okapi BM25 feature weighting. Information retrieval, 14, 466-487.

[13] Torgo, L., & Ribeiro, R. (2009). Precision and recall for regression. In Discovery Science: 12th International Conference, DS 2009, Porto, Portugal, October 3-5, 2009 12 (pp. 332-346). Springer Berlin Heidelberg.

[14] Vujović, Ž. (2021). Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications, 12(6), 599-606.