ISJET

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY

Volume 9 No. 1 January-June 2025

ISSN 2586-8527 (Online)

Panyapiwat Institute of Management

Indexed in the Thai-Journal Citation Index (TCI 1)

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)

Volume 9 No. 1 January-June 2025

ISSN 2586-8527 (Online)
PANYAPIWAT INSTITUTE OF MANAGEMENT

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)
Volume 9 No. 1 January-June 2025
ISSN 2586-8527 (Online)

Copyright
Panyapiwat Institute of Management
85/1 Moo 2, Chaengwattana Rd.,
Bang Talat, Pakkred,
Nonthaburi, 11120, Thailand
Tel. +66 2855 1560
E-mail: isjet@pim.ac.th
Website: https://ph02.tci-thaijo.org/index.php/isjet/index

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)

Volume 9 No. 1 January-June 2025 ISSN 2586-8527 (Online)

Objective:
International Scientific Journal of Engineering and Technology will be dedicated to serving as a forum to share knowledge on research advances in all fields of sciences: Engineering, Technology, Innovation, Information Technology, Management Information Systems, Logistics and Transportation, Agricultural Science and Technology, Animal Science and Aquaculture, Food Science, and other areas in Sciences and Technology. Submissions are welcomed from both PIM as well as other Thai and foreign institutions.

Scope:
Engineering, Technology, Innovation, Information Technology, Management Information Systems, Logistics and Transportation, Agricultural Science and Technology, Animal Science and Aquaculture, Food Science, and other areas of Sciences and Technology.

Type of Article:
• Research article
• Academic article
• Review article

Languages of academic works:
An article written in either English language is accepted for publication.

Reviewing Policy:
1. Any manuscript to be accepted for publication must have been reviewed and approved by at least three peer reviewers in that particular field or related fields. The Journal has a double-blind peer review policy which means that neither the peer reviewer nor the author knows the identity of each other.
2. The submitted manuscript must have never been published in any other periodical, and must not be in the approving process for publication by any other periodical. Also, the author must not plagiarize the work of other people.
3. The article, expression, illustrations, and tables that are published in the Journal are the sole responsibility of the author, and definitely not that of Panyapiwat Institute of Management.
4. The Editorial Board of International Scientific Journal of Engineering and Technology reserves the right to change or revise the name(s) and unit(s) of the author(s) in all cases after the issuance of the letter.
5. The Editorial Board of International Scientific Journal of Engineering and Technology reserves the right to cancel the publication that has been issued a certification of publication in the Journal.
6. The Editorial Board of International Scientific Journal of Engineering and Technology reserves the right for decision making on publishing any article in the Journal.

Frequency of Publication:
Twice a year
• The first issue: January-June
• The second issue: July-December

Publication and Access Charges:
There are no charges to submit and publish all types of articles. Full articles in PDF format can be downloaded free from the journal website at https://ph02.tci-thaijo.org/index.php/isjet/index

ISJET Journal Editorial Board
The office of Research and Development
Panyapiwat Institute of Management
85/1 Moo 2, Chaengwattana Rd.,
Bang Talat, Pakkred, Nonthaburi, 11120, Thailand
Tel. +66 2855 1560
E-mail: isjet@pim.ac.th
Website: https://ph02.tci-thaijo.org/index.php/isjet/index

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)
Volume 9 No. 1 January-June 2025
ISSN 2586-8527 (Online)

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)
Volume 9 No. 1 January-June 2025
ISSN 2586-8527 (Online)

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)
Volume 9 No. 1 January-June 2025
ISSN 2586-8527 (Online)

Dear Colleagues,

In the contemporary era, information and communication technology play a critical role in driving economic growth, social development, and transforming the way people live across the globe. The generation of new knowledge through research and academic exchange has become a vital mechanism for fostering innovation and promoting sustainable technological advancement. The journal cordially invites researchers, academics, students, and practitioners to submit original research articles, academic papers, or review articles in the field of information technology and related disciplines for consideration and publication. I would like to take this opportunity to thank everyone who has complemented our goal by contributing to the INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET).

With kind regards,

Asst. Prof. Dr. Nivet Chiravichitchai
Associate Editor of Engineering and Technology
isjet@pim.ac.th

INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET)
Volume 9 No. 1 January-June 2025
ISSN 2586-8527 (Online)

CONTENTS

A Review of Pedestrian Information Retrieval Research

Yan Xie[1] and Jian Qu[2*]
[1,2]Faculty of Engineering and Technology, Panyapiwat Institute of Management,
Nonthaburi, Thailand
E-mail: 6572100049@stu.pim.ac.com, jianqu@pim.ac.com*

Received: January 4, 2024 / Revised: July 5, 2024 / Accepted: July 10, 2024

***Abstract*—Pedestrian Information Search (PIS) has gained attention for its wide range of practical applications. The main objective of PIS is to find a matching object in a set of scene images or videos. Early work on PIS focused on image-based search. With the advent of deep neural networks, PIS can be freed from the limitations of the search source. Therefore, a systematic study of PIS is necessary. In this paper, we review the research results of PIS based on different modalities in terms of the origin of the PIS task, the development history of PIS, and the methods of training and evaluation of PIS models. We selected the better-performing models for experiments. We summarize and comparatively evaluate the experimental results. Finally, we discuss some of the present problems of PIS and some meaningful future research directions.**

***Index Terms*—Pedestrian Information Search, Pedestrian Detection, Pedestrian Re-Identification, Deep Learning, Neural Network Models**

I. Introduction

Pedestrian Information Search (PIS) is an important and challenging task in computer vision, especially in human-targeted tasks. PIS aims to achieve an effective search for target pedestrian information in a variety of search scenarios. PIS has great potential for application in real-world search scenarios of surveillance videos. Therefore, this paper presents a comprehensive survey of work related to PIS.

PIS is an end-to-end technique for Pedestrian Detection (PD) [1] and Pedestrian Re-Identification (PReI) [2] in panoramic images. PIS needs to accurately derive the coordinate position information and identity information of pedestrians in the image. Considering the actual pedestrian information retrieval function, the PIS can be divided into the joint execution of PD and PReI tasks.

Since 2004, PD has received extensive attention and research [3]. PD, as a kind of target detection, mainly extracts features by manually designed feature extraction methods [4]. For example, the study in [5] achieved PD by designing a unified framework approach. However, the research of Enzweiler and Gavrila [6] showed that the accuracy of manual feature extraction is not high. Meanwhile, the emergence of deep neural networks has brought a new development direction for PD technology. Liu and Stathaki used CNN networks to complete the detection task in their research [7]. The research results of Zhai *et al.* [8] show that deep neural networks can greatly improve accuracy.

PReI was proposed for this task as early as 1996 [9]. However, PReI gained widespread attention after being reintroduced in 2006 at the CVPR (International Conference on Computer Vision and Pattern Recognition). PReI is a technique for determining the presence or absence of specific target pedestrian information in an image or a video sequence [10]. PReI has been reintroduced for two main reasons: 1) the lack of acquisition of pedestrian information in the research conducted at that time [11] and 2 the deep neural networks have been applied so that higher-level features can be acquired [12]. PReI can acquire more and deeper pedestrian information. In PD systems, pedestrian sample information is always ignored. Therefore, better accuracy and efficiency can be achieved by linking PD and PReI tasks [13]. Researchers have indicated that the combination of PD and PReI techniques can indeed enhance the performance of pedestrian information retrieval [14]. Unlike PD and PReI, the main challenge of PIS is to query the gap between people. PIS needs to deal with extra details. PIS can be categorized into Image-based Pedestrian Information Search (IPIS) and Natural Language-based Pedestrian Information Search (NLPIS) according to the search source.

Image-based Pedestrian Information Search (IPIS) is performed using the detection image as the search source. IPIS benefits from the fact that there is the source for conducting the search is explicit. In the study of Sun *et al.* [15], they successfully implemented pedestrian information retrieval using images by training a CNN model. However, IPIS suffers from the limitation of search sources [15]. IPIS cannot be applied in many scenarios. When the search source image is not available, free-form natural language-based character search is very convenient [16].

NLPIS is another major PIS category that uses free-form natural language as a search query. NLPIS is more challenging than the IPIS problem. In practice, the effects of factors such as morphology, occlusion, resolution, and background responsibility can make the PIS task more challenging [17]. Therefore, NLPIS requires that discriminative features need to be learned first before text character matching.

The feature extraction methods of NLPIS are mainly divided into manual feature-based methods and deep feature-based methods. For handcrafted feature-based pedestrian information retrieval methods, a common approach will use handcrafted features such as Histogram of Oriented Gradients (HOG), Local Binary Patterns (LBP), and Color Histograms [18] to obtain information such as shapes, colors, and textures. Hand-designed features are usually interpretable and understandable because they are constructed based on domain expertise. However, hand-designed features are relatively sensitive to changes and deformations in the data and may be less adaptable to changes in lighting, viewing angle, and background. For deep feature-based approaches, use deep learning models to learn higher-level feature representations from images. The deep features can be selected from the deep learning models suitable for pedestrian information retrieval. Deep features are acquired with more freedom compared to manual features [19]. Deep neural networks commonly used for deep features are Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN) [20], etc. The application of deep neural networks also foresees the need for data preprocessing and training of deep neural network models. The pre-processing workload for deep feature extraction is relatively high.

While manual feature-based methods are still effective in some scenarios [21], manual feature-based pedestrian information retrieval methods do not perform well when dealing with complex scenarios and highly variable conditions [22]. Manually designed features cannot capture all the variations and information in the data. Deep learning models can automatically learn the features in the data to better adapt to different scenes and changes. The rise of deep learning methods in recent years has gradually replaced some of the manual feature-based methods. The advantages of deep learning methods are more significant when dealing with large-scale and complex datasets [23]. Overall, deep feature-based PIS methods have achieved significant results in many applications [24].

NLPIS will also involve the matching relationship between two modalities of information, text and image. In the early stage of research on PIS, PIS used unimodal learning. Single modality learning in PIS refers to the process in which the model learns information from image perception modalities only [25]. Single-modality learning is relatively simpler and intuitive because the model only has to process data from one perceptual modality. Therefore, single modality learning is computationally efficient and has a relatively single task. With the improvement of PIS requirements, PIS is no longer a simple determination of whether there is a human or not. Therefore, the necessity of cross-modal pedestrian information retrieval has been emphasized in more studies [26].

Cross-modal refers to the process of information interaction or learning between different perceptual modalities. The cross-modal approach allows the system to acquire rich information from different perceptual modalities, which helps to improve the comprehension and representation of the input data by the system. Cross-modal learning helps to improve the generalization ability of pedestrian information retrieval models as it can learn more abstract and generic representations from multiple modalities, rather than just representations specific to one modality. Overall, unimodal learning may not be able to handle the complex relationships of multimodal information in real-world applications involving multimodal inputs, thus limiting its use in these scenarios. Single-modality learning is suitable for specific tasks and data contexts but has some limitations in processing multimodal information and improving generalization [27]. With the research and development of multimodal learning, more and more approaches are exploring how to effectively use information from different modalities to improve system performance.

Some relevant datasets and evaluation metrics are also investigated. Common quantitative evaluation metrics in the field of pedestrian information search: Precision, recall, F1 score, rank-k accuracy, and Mean Average Precision (MAP) score. Meanwhile, we selected the latest PIS research projects for replication [28]. We will use rank-k accuracy and mAP to evaluate the existing models.

II. Related Studies

In this section, we introduced the tasks associated with the PIS task.

A. Pedestrian Detection

Pedestrian Detection (PD) is the detection of traveling people in the input image. PD needs to locate the position of pedestrians. The application scenarios of PD are extremely wide, including but not limited to pedestrian retrieval in surveillance and automated driving. The main step in the study of PD as a target detection task is to select the region by traversing. Feature extraction is performed during PD with manually designed feature extraction methods. Finally, PD can be achieved by classifying the extracted features. Cao et al. [29] proposed a detection framework that relies on handcrafted features and linear classifiers to achieve PD. Following the 2004 PD, the detectors were improved based on the research of Ribeiro et al. [30], and ICF, ACF, LDCF, and SCF were proposed [31].

However, the development of pedestrian detection methods has also taken a turn for the worse with the emergence of deep neural networks. The emergence of deep neural networks has brought a new direction to pedestrian detection techniques. Researchers have combined artificial features with stronger classifiers. Ribeiro *et al*. [30] used SCF as a detector combined with a deep neural network as a classifier for the detection task. Experiments have shown that access to deep neural networks has improved the accuracy of pedestrian retrieval substantially. Byeon and Kwak [31] used an ACF detector and trained an R-CNN-type neural network to generate pedestrian candidates. The study of Sheng *et al.* [32] implements the subdivision of pedestrian detection into pedestrian attributes and scene attributes [33]. Combine filtered channel features with CNN networks, following the traditional idea of manually designing a feature convolution kernel [34]. Propose an algorithm for learning complexity-aware cascades by seamlessly integrating manual and CNN features into a unified detector. Ma and Gao [35] use LDCF detectors and CNN models to construct part pools for local detection to deal with occlusion problems. Cai *et al.* [34] achieve the best trade-off between accuracy and speed. Meanwhile, [36] shows that deep neural networks can automatically learn high-level features of the target object without relying on manually designed feature extraction methods. Deep neural networks can extract robust features that are independent of the environment, increasing the robustness of detection.

In the early stage of pedestrian detection, researchers use R-CNN to generate candidate suggestions first, and then apply classification and regression algorithms to filter the candidate suggestions [37]. Based on R-CNN, Fast R-CNN, proposed by Zhang *et al.* [38] achieves further improvement in detection time and performance.

B. Pedestrian Re-Identification

Pedestrian Re-Identification (PRI) is also known as Pedestrian Re-Identification. PRI is a technique for determining the presence or absence of a target pedestrian in an image or video sequence [10]. The current research direction of the PRI technique can be roughly divided into feature extraction and metric learning.

In terms of feature extraction, most of the PRI research uses a combination of manual features and deep features. PRI researchers will first extract distinguishable features using manually designed feature extraction, and then learn higher-level features through deep learning neural networks. More PRIs innovate in structure to improve performance. For example, [39] designed two new convolutional layers to obtain the relationship between pairs of pedestrian images whose inputs have been aligned and cropped. Chen *et al.* [40] designed four convolutional layers and 2 fully connected layers to extract feature information from pedestrian images. In metric learning, PRI solves the problem of PRI by learning a distance metric [41]. Proposed KISSME, which uses likelihood ratios to determine similarity using statistical inference, proposed the null space to solve the problem of small sample sizes encountered in metric learning [42].

Traditional deep learning methods mainly use pairwise or ternary distance loss functions to supervise the training process [43], [39] input a pair of cropped pedestrian images into the network. Utilized ternary samples and managed to make the feature distances between pedestrian samples of the same identity as close as possible. Another approach is to consider the PRI problem as a multiclassification problem. Similarly, as the number of categories increases, using the Softmax loss function for PRI makes the process very slow or even fails to converge [44].

From the development history of PRI, we can find that after 2014, deep learning-based PRI task models have gradually gained the favor of most researchers. However, due to the size and singularity of the dataset, deep learning has not achieved as much success in PRI as other computer vision techniques. There is still much space for improving the performance and scene applicability of PRI tasks.

C. Pedestrian Information Search

Pedestrian Information Retrieval (PIS) is an end-to-end technique for detecting and recognizing pedestrians in panoramic images. PIS outputs information about the coordinates of the position of pedestrians in the image, as well as information about their identity. In terms of actual functionality, PIS can be considered as a joint task of PD and PRI. However, simply connecting the two tasks together cannot obtain good accuracy and efficiency.

Xu *et al.* [45] achieved PIS by modeling the common and unique characteristics of pedestrians through a sliding window search strategy. However, their research results show that simply connecting the two tasks cannot achieve good accuracy and efficiency. Proposed an end-to-end single CNN PIS framework [46]. Xiao *et al.* achieved simultaneous processing of two tasks in a single CNN. Meanwhile, proposed an Online Instance Matching (OIM) loss function to effectively improve the performance of neural networks. Investigated the overall impact of the performance of the pedestrian detection component in the pedestrian search task. Used a two-stage strategy to implement PIS with a pedestrian detection network cascaded with a pedestrian re-identification network [46], [47]. Liu *et al.* [48] used a recurrent neural network to correct the pedestrian position in the panoramic image step by step and match the pedestrians. In traditional PIS studies, the dataset only contains manually cropped pedestrian frames. Contributed a new large-scale dataset, PRW, for pedestrian search. These works are aimed at making PIS applicable in integrated scenarios. These works aim to incorporate PD and PRI into a complete framework to reduce the mutual influence of the errors of the two otherwise independent networks [49].

The proposed PIS has made its application scenarios to become more leading with higher market requirements for PIS. PIS in severe non-aligned scenarios is a typical scenario. The focus of severe non-aligned scenarios is to utilize multiple features of the pedestrian images to achieve a reliable search of target pedestrians. Specifically, the facial features of the target will first be used to expand the target pedestrian samples to indirectly search for target pedestrians with large differences in body shape and appearance in the image. Then, the search results will be used to reverse search the sample with face information, and the similarity between the face features of the sample and the target face features will be used to filter the search results to obtain the final search results. The non-aligned scenario is closer to the actual application scenario, and the research is very significant. For example, in the study of Zheng, Gong, and Xiang, they proposed a Probabilistic Relative Distance Comparison (PRDC) model to reduce the distance between true matches and false matches [50]. Used the Deformable Part Model (DPM) to generate the Market-1501 dataset. Zheng *et al.* proposed a BoW descriptor method to try to bridge the gap between image searches [51]. Sun et al. proposed a Part-based Convolutional Baseline (PCB) to learn the features noted by parts. The PCB uses a simple uniform partitioning method to assemble some of the informative features into convolutional descriptors. However, their research requires images that have been used as a search source. It also makes PIS limited in many scenarios. Therefore, free natural language-based PIS is being seen in more and more studies [52].

Li *et al.* [16] proposed a recurrent neural network (GNA-RNN) with a gated neural attention mechanism using a recurrent neural network structure to solve the problem of affinity between textual descriptions and images of people. Krizhevsky *et al.* [53] used 1.3 million high-resolution images from the LSVRC-2010 ImageNet training set to train a large deep convolutional neural network to implement image-based PIS [54]. Modified the pre-trained BERT network by introducing Sentence-BERT (SBERT). BERT is a well-trained network for NLP [55], [56]. Reimers et al implemented text-based PIS by using concatenated and ternary network structures. However, the effect of unimodal PIS is limited. Then, a cross-modality-based feature extraction approach was verified to enhance the performance of PIS [57]. Multiple feature extraction requires multiple types of information to be concentrated in a single system [28]. Proposed an end-to-end learning framework, TIPCB. Using a multi-stage cross-modal matching approach, visual and textual representations are matched at multiple levels. Zhang *et al.* [58] proposed two losses, Cross-Modal Projection Matching (CMPM) and Cross-Modal Projection Classification (CMPC), which achieve image-text cross-modal feature extraction [57]. Proposed an end-to-end Simple and Robust Correlation Filtering (SRCF) framework to extract key information and adaptively align local features without the need for auxiliary tools. In summary, multimodal-based PIS performs better than unimodal PIS. A schematic of a typical cross-modal PIS model is shown in Fig. 1.

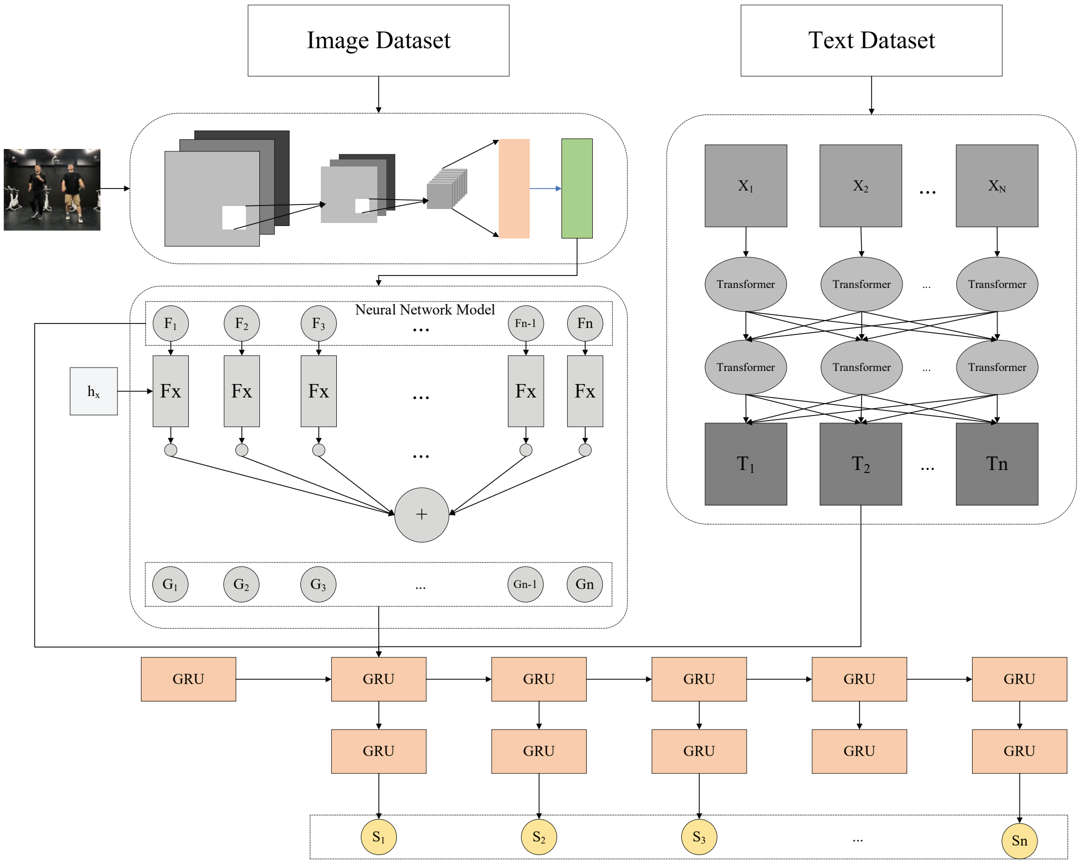

Fig. 1. A typical deep neural network model for cross-modal pedestrian information retrieval

Meanwhile, the research of Ye *et al.* [57] shows that a suitable sample set can effectively improve the performance of PIS. Existing PIS large-scale image datasets that are heavily used are the MS COCO dataset, the ImageNet dataset, and the CUHK-PEDES dataset. The MS COCO dataset was first created and released by Microsoft Research in September 2014. MS COCO, as a large-scale image dataset designed for the task of computer vision, has become one of the most important benchmarks in the field of computer vision. One of the important benchmarks in the field of computer vision. The data in the MS COCO dataset is rich and diverse. The richness and diversity of MS COCO make it an ideal data source for computer vision tasks. At the same time, the MS COCO dataset has gone through several updates to meet the evolving needs of computer vision research. The MS COCO dataset has been progressively added with more images, detailed annotations, and support for different tasks. The MS COCO dataset updating process not only increases the size of the dataset, but also improves the coverage of the dataset in terms of complex scenarios and a wide variety of objects. The MS COCO data offers researchers have more challenging and varied data sources for computer vision tasks. researchers with more challenging and practically relevant data support [59].

The ImageNet dataset is a large-scale image database created by Stanford University [60]. The ImageNet dataset is designed to facilitate research in the field of computer vision. The ImageNet dataset contains more than 14 million images. The images in the ImageNet dataset cover more than 20,000 different categories of objects. The categories of the ImageNet dataset cover a wide variety of objects, ranging from animals and plants to everyday objects. The image annotation of the ImageNet dataset is the addition of relevant labels and annotations to each image. The categories of the ImageNet dataset cover a wide range of objects, from animals and plants to everyday objects. The image annotation of the ImageNet dataset is the addition of relevant labels and comments to each image. The labels of the ImageNet dataset describe the main objects or scenes that appear in the image. The comprehensiveness and the wide range of applications of the ImageNet dataset have made it an important part of the image classification task.

The CUHK-PEDES dataset was created and released by the Research Institute of the Chinese University of Hong Kong [16]. The CUHK-PEDES dataset is a rich pedestrian dataset with annotations. The CUHK-PEDES dataset is the first dataset created specifically for PIS. The CUHK-PEDES dataset aggregates images from five existing pedestrian re-identification datasets, including CUHK03, Market-1501, SSM, VIPER, and CUHK01, resulting in a large dataset of 40,206 images containing more than 13,003 individuals. Each image in the CUHK-PEDES dataset has been carefully annotated by staff [16]. CUHK-PEDES uses two textual descriptions to provide exhaustive details on the appearance, movements, and poses of the characters. The textual descriptions are rich in information, making the dataset more challenging and complex for practical applications [61]. Overall, all the above datasets play an important role in computer vision research. Different datasets provide rich and extensive data resources for model training and evaluation for image classification, target detection, and other related tasks.

The PIS research process is shown in Fig. 2. All the experiments were conducted based on this training-validation-testing division, which helps the researcher to compare and evaluate in a standard experimental setup. For performance evaluation, the researcher uses different metrics to evaluate the performance of the PIS task. The PIS task acts as a retrieval task. For a given target pedestrian image, after comparing it with the features of the samples in the candidate set, the similarity between the query pedestrian image and all the samples in the candidate set is calculated, and finally, all the candidate targets are sorted according to the similarity from highest to lowest. When the similarity between the query image and the matching image is higher, it means that the performance of the pedestrian search model is better. In the field of PIS, the commonly used quantitative evaluation metrics are Rank-K accuracy [62] and mean Average Precision (mAP) score [63]. In our study, we used Top1, Top5, and Top10 in the training model to evaluate the performance of the model. mAP is the average of the mean accuracy of all samples in the query set. This division provides researchers with a balanced and diverse dataset for model training and performance evaluation.

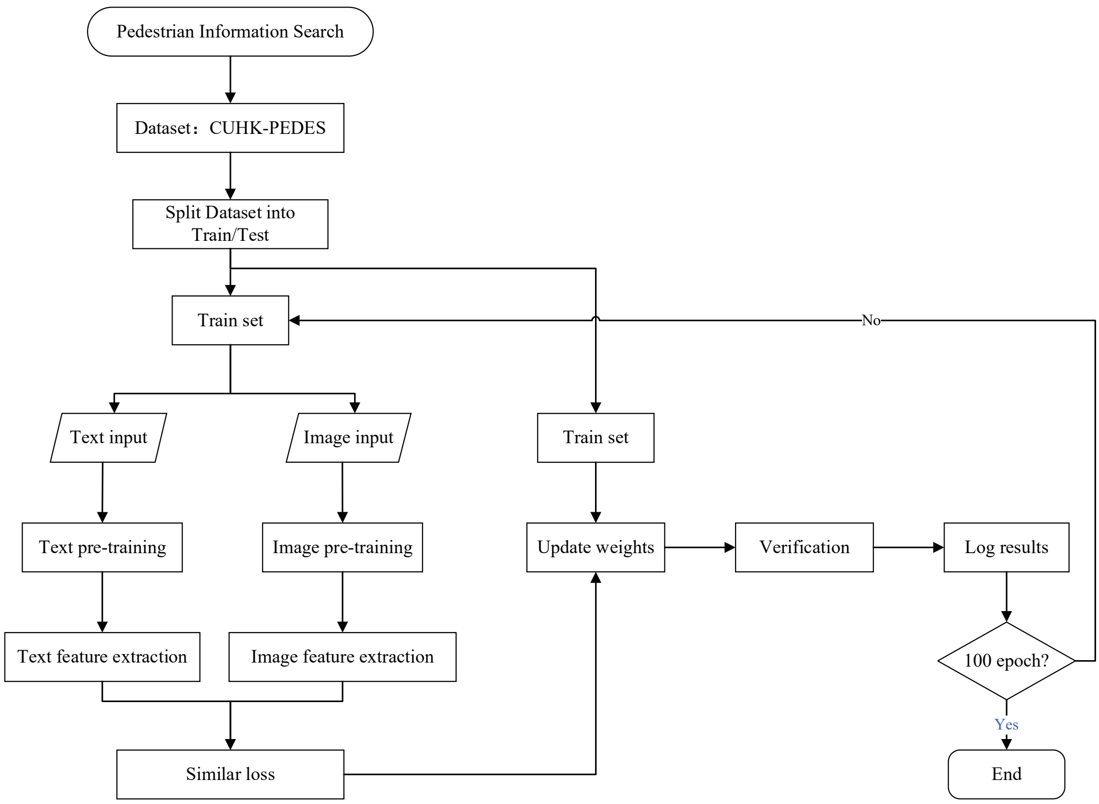

Fig. 2. Training process of a deep neural network model for cross-modal pedestrian information retrieval

III. Comparing Model Methods

In this section, we have selected different models and datasets for the training of PIS models. For the PIS task, the main choices are the selection of the PIS model and the selection of the training dataset. For model selection, we chose the cross-modal pedestrian information retrieval model, the Res50Bert model. Res50Bert model is a combination of ResNet50 and bert-base-uncased. Res50Bert model will train both natural language feature information and image feature information to improve the performance of PIS. Meanwhile, the performance of the Res50Bert model is better than the performance of other models in existing PIS research [28]. Therefore, we choose the Res50Bert model as the basic model for our validation. For dataset selection, we chose the

CUHK-PEDES dataset as the dataset designed for the PD task. The detailed description of pedestrians in the CUHK-PEDES dataset helps the model to better understand the semantic information in the images, which in turn improves its performance in real scenarios. The CUHK-PEDES dataset is divided into three non-overlapping subsets, which are used for training, validation, and testing, ensuring that individuals with the same identity do not appear in different sets. In addition, for performance evaluation, we adopted top-k accuracy as the primary metric for the person retrieval task. At the same time, we also calculated the mAP to analyze the overall performance of the model. Meanwhile, the study of Suo et al. [64] shows that the CUHK-PEDES dataset has better PIS performance. Therefore, we choose the CUHK-PEDES dataset as the training dataset.

In the process of studying the PIS task, we found that among the image neural network models, the performance of the EfficientNet neural network is improved in both efficiency and accuracy compared to the ResNet neural network model. Meanwhile, in the process of understanding text models, we also found several text models applicable to pedestrian information retrieval: paraphrase-multilingual-MiniLM -L12-v2 [65], distiluse-base-multilingual-cased-v2 [66], bert-base-nli-mean-tokens [67], all-mpnet-base-v2 [68], MiniLM-L12-H384-uncased [69]. Thus, we trained the base model and its different combinations as shown in Table I.

TABLE I
CROSS-MODAL PIS MODEL TRAINING RESULTS

Pedestrian Information Retrieval Model	Image Model	Text Model	Rank1	Rank5	Rank10	mAP
Res50Bert	ResNet50	best-base-uncased	0.595992	0.800582	0.867647	0.507627
Res50PMML12V2	ResNet50	paraphrase-multilingual-MiniL M-L 12-v2	0.577085	0.789754	0.860698	0.493230
Res50DBMCV2	ResNet50	distiluse-base-multilingual-cased-v2	0.584034	0.792178	0.863445	0.497706
Res50BBNMT	ResNet50	bert-base-nli-mean-tokens	0.573368	0.791370	0.864092	0.490792
Res50AMBV2	ResNet50	all-mpnet-base-v2	0.588235	0.789916	0.868778	0.500985
Res50MLH384U	ResNet50	MiniL M-L 12-H384-uncased	0.578539	0.786316	0.837750	0.497329
EB1Bert	EfficientNet B1	best-base-uncased	0.605992	0.805582	0.869647	0.509627
EB1PMML12V2	EfficientNet B1	paraphrase-multilingual-MiniL M-L 12-v2	0.314156	0.554299	0.665482	0.264532
EB1DBMCV2	EfficientNet B1	distiluse-base-multilingual-cased-v2	0.526503	0.754848	0.840175	0.450074
EB1BBNMT	EfficientNet B1	bert-base-nli-mean-tokens	0.499199	0.742400	0.826811	0.425028
EBI AMBV2	EfficientNet B1	all-mpnet-beste-v2	0.577085	0.789754	0.860698	0.493230
EBIMLH384U	EfficientNet B1	MiniL M-L 12-H384-uncased	0.080478	0.214447	0.316742	0.076245

From Table I, we found that the basic model reaches about 60%. The experimental results indicate that both different text models and image models affect the performance of the PIS model. At the same time, the experimental results also illustrate that there is a possibility that the combination of different text models and image models can enhance the performance of PIS.

IV. CONCLUSION AND FUTURE WORK

Our study focuses on a systematic review of pedestrian information retrieval. Firstly, we introduce the pedestrian search task and the pedestrian re-identification task. After the introduction of pedestrian search and pedestrian re-identification focus is on the pedestrian information retrieval task, which is a combination of the two tasks. In this paper, we focus on the feature extraction method, modality, and dataset of PIS. We summarize the performance of unimodal and multimodal pedestrian information retrieval tasks. Existing research shows that multimodality is fully capable of implementing free-form natural language-based PIS. The implementation of natural language-based pedestrian information retrieval also heralds the possibility of a higher degree of freedom in human-computer interaction. Meanwhile, we validate the current best-performing model using experimental replication. The experimental results show that although the PIS performance has been improved, the accuracy still cannot reach a high level. Therefore, natural language-based PIS still deserves more time and effort.

At the same time, the existing studies are incomplete. First of all, the existing studies have been conducted in English. There is a lack of research on other languages. Other languages, such as Chinese, which has a large number of speakers in the world, and Thai and Japanese, which are widely spoken, are also small languages that are worth studying. Correspondingly, the existing datasets are all labeled and annotated in English. There are differences in the way languages are expressed and used in practice. It is also meaningful to establish exclusive pedestrian information retrieval datasets for different languages.

Secondly, existing research is realized by strong research teams and companies. The implementation of low-performance hardware to train high-performance models can be an important research direction in the future. We can consider how we can utilize the extreme performance of computers. This will make AI reachable.

References

[1] C. Lin, J. Lu, G. Wang, and J. Zhou, "Graininess-aware deep feature learning for robust pedestrian detection," *IEEE Trans. Image Process.*, vol. 29, pp. 3820-3834, Jan. 2020.

[2] J. Zhang, N. Wang, and L. Zhang, "Multi-shot pedestrian re-identification via sequential decision making," in *Proc. IEEE CVPR*, 2018, pp. 6781-6789.

[3] C. J. Pai, H. R. Tyan, Y. M. Liang, and H. Y. Mark Liao, "Pedestrian detection and tracking at crossroads," *Pattern Recognit.*, vol. 37, no. 5, pp. 1025-1034, Jan. 2003.

[4] M. You, Y. Zhang, C. Shen, and X. Zhang, "An extended filtered channel framework for pedestrian detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1640-1651, May 2018.

[5] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 4, pp. 743-761, Apr. 2012.

[6] M. Enzweiler and D. M. Gavrila, "Monocular pedestrian detection: Survey and experiments," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 12, pp. 2179-2195, Oct. 2009.

[7] T. Liu and T. Stathaki, "Faster R-CNN for robust pedestrian detection using semantic segmentation network," *Front. Neurorobot.*, vol. 64, no. 12, pp. 1-10, Oct. 2018.

[8] S. Zhai, S. Dong, D. Shang, and S. Wang, "An improved Faster R-CNN pedestrian detection algorithm based on feature fusion and context analysis," *IEEE Access*, vol. 8, pp. 138117-138128, Jul. 2020.

[9] C. Wren, A. Azarbayejani, T. Darrell, and A. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780-785, Jul. 1997.

[10] L. Zheng et al. "Mars: A video benchmark for large-scale person re-identification," *Computer Vision-ECCV*, vol. 9910, pp. 868-884, Sep. 2016.

[11] D. Figueira, M. Taiana, A. Nambiar, J. Nascimento, and A. Bernardino, "The HDA+ dataset for research on fully automated re-identification systems," Cham, CH: Springer, 2015, pp. 241-255.

[12] Y. Shen, H. Li, S. Yi, D. Chen, and X. Wang, "Person re-identification with deep similarity-guided graph neural network," in *Proc. ECCV*, 2018, pp. 486-504.

[13] S. Salehian, P. Sebastian, and A. B. Sayuti, "Framework for pedestrian detection, tracking and re-identification in video surveillance system," in *Proc. IEEE ICSIPA*, 2019, pp. 192-197.

[14] S. Zhang, D. Chen, J. Yang, and B. Schiele, "Guided attention in CNNs for occluded pedestrian detection and re-identification," *Int. J. Comput. Vis.*, vol. 129, pp. 1875-1892, Apr. 2021.

[15] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," *ECCV*, vol. 11208, pp. 480-496, Oct. 2018.

[16] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang, "Person search with natural language description," in *Proc. IEEE CVPR*, 2017, pp. 5187-5196.

[17] X. Han, S. He, L. Zhang, Q. Ye, and J. Sun, "Text-based person search with limited data," in *Proc. ACM Int. Conf. Multimedia*, 2023, pp. 1-20.

[18] D. K. Panda and S. Meher, "Dynamic background subtraction using local binary pattern and histogram of oriented gradients," in *Proc. ICIIP*, 2015, pp. 306-311.

[19] C. Wang, Z. Luo, Y. Lin, and S. Li, "Text-based person search via multi-granularity embedding learning," in *Proc. IJCAI*, 2021, pp. 1068-1074.

[20] H. Liu, J. Feng, M. Qi, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492-3506, Jul. 2017.

[21] P. Srivastava and A. Khare, "Utilizing multiscale local binary pattern for content-based image retrieval," *Multimedia Tools Appl.*, vol. 77, pp. 12377-12403, Jun. 2017.

[22] K. Chen, X. Song, X. Zhai, B. Zhang, B. Hou, and Y. Wang, "An integrated deep learning framework for occluded pedestrian tracking," *IEEE Access*, vol. 7, pp. 26060-26072, Feb. 2019.

[23] D. Jiang and M. Ye, "Cross-modal implicit relation reasoning and aligning for text-to-image person retrieval," in *Proc. IEEE/CVF*, 2023, pp. 2787-2797.

[24] J. S. J. Rani and M. G. Augasta, "PoolNet deep feature based person re-identification," *Multimedia Tools Appl.*, vol. 82, no. 16, pp. 24967-24989, Jan. 2023.

[25] L. Zheng, Y. Yang, and A. G. Hauptmann, "Person re-identification: Past, present and future," *J. LaTeX Class. File*, vol. 14, no. 8, pp. 1-20, Aug. 2016.

[26] Z. Wang, Z. Wang, Y. Zheng, and Y. Y. Chuang, "Learning to reduce dual-level discrepancy for infrared-visible person re-identification," in *Proc. IEEE/CVF CVPR*, 2020, pp. 618-626.

[27] Y. Sun, L. Zheng, W. Deng, and S. Wang., "SVDNet for Pedestrian Retrieval," in *Proc. 2017 IEEE ICCV*, 2017, pp. 3820-3828.

[28] Y. Chen, G. Zhang, Y. Lu, Z. Wang, Y. Zheng, and R. Wang "TIPCB: A simple but effective part-based convolutional baseline for text-based person search," *Neurocomputing*, vol. 494, pp. 171-181, 2002.

[29] J. Cao, Y. Pang, J. Xie, F. S. Khan, and L. Shao, "From handcrafted to deep features for pedestrian detection: A survey," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 44, no. 9, pp. 4913-4934, Apr. 2021.

[30] D. Ribeiro, J. C. Nascimento, A. Bernardino, and G. Carneiro, "Improving the performance of pedestrian detectors using convolutional learning," *Pattern Recognit*., vol. 61, pp. 641-649, 2017.

[31] Y. H. Byeon and K. C. Kwak, "A performance comparison of pedestrian detection using faster RCNN and ACF," in *Proc. 2017 6th IIAI Int. Congr. Adv. Appl. Inform.*, 2017, pp. 858-863. https://doi.org./10.1109/IIAI-AAI.2017.196

[32] B. Sheng, Q Hu, J. Li, W. Yang, B. Zhang, and C. Sun, "Filtered shallow-deep feature channels for pedestrian detection," *Neurocomputing,*vol. 9, pp. 106-113, Aug. 2017.

[33] S. Zhang, R. Benenson, and B. Schiele, "Filtered feature channels for pedestrian detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1751-1760.

[34] Z. Cai, M. Saberian, and N. Vasconcelos, "Learning complexity-aware cascades for deep pedestrian detection," in *Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 3361-3369.

[35] Z. Ma and P. P. Gao, "Research on the Cascade Pedestrian Detection Model Based on LDCF and CNN," in *Proc. 2018 IEEE Int. Conf. Big Data Smart Comput. (BigComp)*, 2018, pp. 314-320.

[36] Y. Luo, C. Zhang, M. Zhao, H. Zhou, and J. Sun "Where, what, whether: Multi-modal learning meets pedestrian detection," *Computer Vision Foundation,* no. zrxiv2103.11599, pp. 14065, Dec. 2020.

[37] P. Dong and W. Wang, "Better region proposals for pedestrian detection with R-CNN," in *Proc. 2016 Visual Commun. Image Process. (VCIP)*, 2017, pp. 1-4.

[38] H. Zhang et al., "Pedestrian Detection Method Based on Faster R-CNN," in *Proc. 2017 13th Int. Conf. Comput. Intell. Security (CIS)*, 2018, pp. 427-430.

[39] E. Ahmed, M. Jones, and T. K. Marks, "An improved deep learning architecture for person re-identification," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 3908-3916.
[40] X. Chen, Y. Xu, D. W. K. Wong, T. Y. Wong, and J. Liu, "Glaucoma detection based on deep convolutional neural network," in *Proc. 2015 37th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC*), 2015, pp. 715-718.
[41] M. Koestinger, M. Hirzer, P. Wohlhart, P. M. Roth, and H. Bischof, "Large scale metric learning from equivalence constraints," in *Proc. 2012 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2012, pp. 2288-2295.
[42] L. Zhang, T. Xiang, and S. Gong, "Learning a discriminative null space for person re-identification," in *Proc. 2016 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 1239-1248.
[43] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037-3045, Oct. 2019.
[44] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. 2015 IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 815-823.
[45] Y. Xu, X. Cao, and H. Qiao, "An efficient tree classifier ensemble-based approach for pedestrian detection," *IEEE Trans. Syst., Man, Cybern. B, Cybern*., vol. 41, no. 1, pp. 107-117, May 2011.
[46] T. Xiao, S. Li, B. Wang, L. Lin, and X. Wang, "Joint detection and identification feature learning for person search," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 3415-3424.
[47] H. Liu, J. Feng, M. Qi, J. Jiang, and S. Yan, "End-to-end comparative attention networks for person re-identification," *IEEE Trans. Image Process.*, vol. 26, no. 7, pp. 3492-3506, May 2017.
[48] L. Zheng, H. Zhang, S. Sun, M. Chandraker, Y. Yang, and Q. Tian, "Person re-identification in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 3346-3355.
[49] W. S. Zheng, S. Gong, and T. Xiang, "Person re-identification by probabilistic relative distance comparison," in *Proc. CVPR 2011*, 2011, pp. 649-656.
[50] L. Zheng, L. Shen, L. Tian, J. Wang, and Q. Tian "Scalable person re-identification: A benchmark," in *Proc. 2015 IEEE Int. Conf. Comput. Vis. (ICCV)*, 2015, pp. 1116-1124.
[51] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, 2017, pp. 480-496.
[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84-90, Dec. 2012.
[53] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proc. EMNLP-IJCNLP*, 2019, pp. 3982-3992
[54] Q. Li and J. Qu, "A novel BNB-NO-BK method for detecting fraudulent crowdfunding projects," *SJST*, vol. 44, no. 5, pp. 1209-1219, Oct. 2022.
[55] W. Hou and J. Qu, "BM5-SP-SC: A dual model architecture for contradiction detection on crowdfunding projects," *CAST*, vol. 23, no. 6, pp. 1-29, Apr. 2023.
[56] W. Suo et al., "A simple and robust correlation filtering method for text-based person search," *Eur. Conf. Comput. Vis*., vol. 13695, pp. 726-742, Nov. 2022.
[57] Y. Zhang and H. Lu, "Deep cross-modal projection learning for image-text matching," *Springer*, vol. 11205, pp. 686-701, Oct. 2018.
[58] M. Ye, J. Shen, G. Lin, T. Xiang, L. Shao, and S. C. H. Hoi, "Deep learning for person re-identification: A survey and outlook," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 44, no. 6, pp. 2872-2893, Jan. 2021.
[59] T. Y. Lin et al., "Microsoft coco: Common objects in context," *Springer*, vol. 8693, pp. 740-755, Sep. 2014.
[60] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. 2009 IEEE Conf. Comput. Vis. Pattern Recognit*., 2009, pp. 248-255.
[61] X. H. Phan, L. M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 91-100.
[62] Y. Zhang, R. Alturki, H. J. Alyamani, and M. Ikram, "Multilabel CNN-based hybrid learning metric for pedestrian reidentification," *Mobile Inf. Syst*., vol. 2021, no. 7, pp. 1-7, Apr. 2021.
[63] J. Revaud, J. Almazán, R. S. Rezende, and C. R. de Souza, "Learning with average precision: Training image retrieval with a listwise loss," in *Proc. 2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV*), 2019, pp. 5106-5115.
[64] R. Pan, J. A. García-Díaz, M. Á. Rodríguez-García, and R. Valencia-García, "UMUTeam at SemEval-2023 Task 3: Multilingual transformer-based model for detecting the genre, the framing, and the persuasion techniques in online news," in *Proc. 17th Int. Workshop Semantic Eval*., 2023, pp. 609-615.
[65] B. Bharathi and G. U. Samyuktha, "Machine learning based approach for sentiment analysis on multilingual code-mixing text," in *Proc. FIRE*, 2021. pp. T6-T18.
[66] K. Peyton and S. Unnikrishnan, "A comparison of chatbot platforms with the state-of-the-art sentence BERT for answering online student FAQs," *Results Eng*., vol. 17, p. 100856, Mar. 2023.
[67] G. Ashqar and A. Mutlu, "A Comparative assessment of various embeddings for keyword extraction," in Proc. *5th Int. Congr. Human-Computer Interaction,* 2023, pp. 1-6.
[68] R. Qin, "Bert-based feature extraction approach for software forum posts," *IEEE Access*, vol. 11, pp. 1-9, Jan. 2024, https://doi.org./10.1109/ACCESS.2024.3426976

Yan Xie is currently studying for the Master of Engineering Technology, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand. She received B.B.A from Nanjing Tech University Pujiang Institute, China, in 2022. Her research interests are Research direction is artificial intelligence, image processing, and Natural Language Processing (NLP).

Jian Qu is an Assistant professor at the Faculty of Engineering and Technology, Panyapiwat Institute of Management, Thailand He received a Ph.D. with an Outstanding Performance award from Japan Advanced Institute of Science and Technology, Japan, in 2013. He received B.B.A with Summa Cum Laude honors from Institute of International Studies of Ramkhamhaeng University, Thailand, in 2006, and M.S.I.T from Sirindhorn International Institute of Technology, Thammast University, Thailand, in 2010. He has been a house committee for Thai SuperAI since 2020. His research interests are natural language processing, intelligent algorithms, machine learning, machine translation, information retrieval, and image processing.

Defect Reduction in Automotive Seat Manufacturing: A Lean Six Sigma Approach

Bundit Wongthong[1], Poom Jatunitanon[2], Bundit Inseemeesak[3], and Yodnapha Ketmuang[4*]
[1]Department of Industrial Engineering, Faculty of Engineering, Thonburi University, Bangkok, Thailand
[2,3,4]Automotive Manufacturing Engineering, Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand
E-mail: bundid.w@gmail.com, poomjat@pim.ac.th, banditins@pim.ac.th, yodnaphaket@pim.ac.th*

Received: April 3, 2025 / Revised: May 19, 2025 / Accepted: May 19, 2025

Abstract— **This study investigates the issue of rear cushion wrinkling in a pickup truck seat production line using the Lean Six Sigma (LSS) methodology. By applying the DMAIC framework, we identified that excessive tension in the extruded listing fleece caused deformation, particularly in curved seat sections. To resolve this problem, we redesigned the fleece by incorporating rectangular slots (15 × 5 mm) spaced 80 mm apart. As a result, wrinkling defects were reduced by 60%, from 312 to 187 pieces, lowering the overall defect rate from 4.05% to 1.61% over six months. This exceeded our initial goal of reducing defects to less than 2.0%. Additionally, this improvement led to estimated cost savings of 852,500 THB, primarily due to a reduction in rework and material waste. Beyond cost benefits, the new design helped streamline the production process, cutting cycle time by 20% and improving customer satisfaction by a similar percentage. While these results demonstrate the effectiveness of Lean Six Sigma in quality improvement, certain limitations remain. Factors such as operator variability and material inconsistencies were not fully controlled in this study. Future research could explore real-time defect detection systems or adaptive tension control mechanisms to enhance process stability.**

Index Terms— **Lean Six Sigma, DMAIC, Wrinkle Defects, Defect Reduction, Automotive Manufacturing**

I. Introduction

In the automotive manufacturing industry, balancing high product quality with efficient production processes is a constant challenge. To address inefficiencies and maintain competitiveness, we applied Lean Six Sigma (LSS), a methodology that combines Lean manufacturing's principles of waste reduction with Six Sigma's emphasis on reducing process variability. Using the DMAIC (Define, Measure, Analyze, Improve, and Control)

Framework, we systematically identified key process inefficiencies in our production line, particularly those affecting defect rates and cycle time. For instance, variations in material handling and assembly techniques contributed to inconsistent quality levels. By integrating data-driven analysis with Lean methodologies, we were able to pinpoint root causes and implement targeted improvements that enhanced overall production efficiency [1].

Previous studies have highlighted the effectiveness of LSS in automotive production, with studies reporting defect reductions of up to 50% and cost savings exceeding 40% in similar projects [2]. In this study, we applied the LSS DMAIC framework to address a critical quality issue—rear cushion wrinkling—in a specific production line at a sample company in the automotive industry, focusing on a seat model for a pickup truck. The wrinkling issue, primarily caused by excessive tension in the listing fleece at the seat's curved sections, not only compromised aesthetic quality but also increased rework costs and impacted customer satisfaction.

Our objective was to systematically investigate this issue, identify its root cause, and implement effective corrective actions. Several approaches were explored, including modifying the fleece design, refining sewing techniques, and utilizing Statistical Process Control (SPC) tools to monitor and maintain quality standards. We then tested these solutions to determine whether the defect rate could be reduced below 2.0% and whether defect-related costs could be decreased by at least 40%. It is important to note that this study focuses solely on a single production line and specifically on the wrinkling defect in pickup truck seats, without extending to other production lines or defect types.

II. Objective

The objective of this study is to reduce wrinkling defects in the MMTH 4P00 seat production line using Lean Six Sigma methodology.

III. Literature Review

A. Six Sigma Theory and Principles

The implementation of Lean Six Sigma in automotive manufacturing has been extensively studied to improve process efficiency, defect reduction, and quality control. George [3], Lean Six Sigma integrates Lean principles with Six Sigma's statistical tools to enhance production capabilities and eliminate non-value-added activities [2]. Further emphasized the impact of LSS in reducing operational costs and increasing process reliability in automotive component manufacturing [2].

B. DMAIC Methodology

The DMAIC (Define, Measure, Analyze, Improve, and Control) framework is widely used in automotive seat manufacturing to enhance production efficiency and minimize defects. It helps manufacturers systematically analyze issues such as fabric misalignment, stitching errors, and assembly inconsistencies. In practice, the Measure phase involves collecting defect data from seat cushion inspections, allowing engineers to identify common quality concerns. During the Analyze phase, they investigate root causes, such as uneven fabric tension contributing to wrinkles. Research has shown that integrating DMAIC into seat production can lead to measurable improvements, including reduced defect rates, greater process stability, and lower manufacturing costs [4].

C. Related Research Studies

Several studies have investigated improvements in car seat manufacturing quality across different contexts [4], [5]. For example, Tsou and Chen [4] proposed an integrated model combining the Economic Production Quantity (EPQ) framework with Six Sigma's DMAIC methodology. This model aimed to reduce production costs and defect rates in automobile seat assembly lines. In addition, Purba and Sunadi [5] carried out fine feature Quality Function Deployment (QFD) within the automobile seat industry, aligning production approaches with customer necessities to improve seat pleasant and personal pride.

Wrinkle reduction in cloth and leather packages Preceding research on wrinkle reduction in car seats has explored numerous techniques, specifically for leather substances, that are at risk of tension-associated wrinkling in the course of assembly. As an instance, adjusting material tension and applying warmth remedies were recognized as powerful strategies to minimize wrinkles while keeping the sturdiness and appearance of leather-based seats. These techniques informed our approach, guiding us to focus on structural modifications to deal with the wrinkling difficulty. Cloth amendment strategies in the Seat production

In the Thai manufacturing sector, research has focused on enhancing manufacturing performance within the automobile industry. Research like [6] displays how Statistical Procedure Control (SPC) tools, together with manage charts and technique mapping, can effectively reduce defects and hold fine consistency. Moreover, Thai researchers have explored how lean manufacturing concepts, paired with Value Engineering, can help streamline car seat manufacturing with the aid of boosting efficiency and reducing fabric waste [7].

IV. Research Methodology

A. Define Phase

The rear cushion wrinkling issue in the MMTH 4P00 production line, identified as a critical quality defect impacting both efficiency and customer satisfaction, prompted a Lean Six Sigma initiative following data analysis from April to June 2023. Out of 25,408 units produced, 1,028 exhibited wrinkles (4.05% defect rate), with Pareto analysis attributing 30.3% of total defects to this issue, primarily caused by excessive tension in the extruded fleece material, particularly in the seat's curved sections. To address the problem peaking in May 2023, the project aims to reduce defects by at least 50% (lowering the rate to below 2.0%) within six months through structural design modifications, standardized sewing/assembly processes, and tension protocol adjustments. Leveraging the DMAIC framework, the initiative also targets a 40% cost reduction by minimizing material waste and rework expenses, while enhancing product durability and customer satisfaction through improved aesthetics and consistency.

B. Measure Phase

The Measure phase established a defect baseline for the MMTH 4P00 production line, focusing on rear cushion wrinkling and other quality issues. Over three months (April-June 2023), data were systematically collected through visual inspections by trained QC teams using standardized checklists to identify defects such as wrinkles, scratches, and assembly errors. Defective units were tagged, recorded in the digital quality management system, and verified against production logs for accuracy. Based on manufacturing data, 1,028 defects were detected among 25,408 parts, yielding a defect rate of 4.05%. This baseline analysis provided a reference point for identifying critical problem areas and monitoring improvements throughout the production process.

Table I summarizes production and defect data from April to June 2023. In April, 7,745 units were produced with 319 defective units (4.12% defect rate). May saw increased output (9,601 units) but recorded the highest defects at 380 units (3.96% rate). June produced 8,062 units with 329 defects (4.08% rate). Fig. 1, a vertical bar chart, visualizes monthly defect counts (April: 319, May: 380, June: 329) and the cumulative 1,028 defective units. The chart underscores May's peak defect count, indicating potential fluctuations in production conditions or material quality during that period—a trend to be investigated in the Analyze phase [8].

To further analyze the defect profile, the 1,028 defective units were categorized into 22 distinct types during inspections. Fig. 2 presents a horizontal bar chart illustrating the ten most common defects observed in production. "Wrinkled Seat" dominated with 312 cases (30.3% of total defects), followed by "Scratches/Scrapes" (143 units, 13.9%), "Foam Debris Inside" (74 units, 7.2%), "Unusual Testing Noise" (72 units, 7.0%), and "Headrest with Vaseline Stains" (67 units, 6.5%). Additional defects ranked in the top ten were "Incorrect Foam Coverage" (51 units), "Incomplete Label" (47 units), "Loose Wiring" (41 units), "Stain Marks" (38 units), and "Visible Needle Holes" (34 units). Together, these defects accounted for 65% of all reported issues, highlighting 'Wrinkled Seat' as the primary concern requiring corrective action. This analysis provided key benchmarks for identifying root causes and monitoring improvement efforts in subsequent project phases.

TABLE I
PRODUCTION AND DEFECT DATA FOR MMTH 4P00 PRODUCTION LINE (APRIL-JUNE 2023)

Month	Total Production	Defective Units	Defect Rate (%)
April	7,745	319	4.12
May	9,601	380	3.96
June	8,062	329	4.08
Total	25,408	1,028	12.00

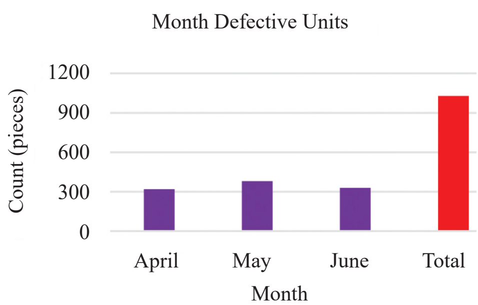

Fig. 1. Vertical bar chart depicting monthly defect counts in the MMTH 4P00 production line (April-June 2023, Total: 1,028 pieces), with May showing the highest defect count at 380 units.

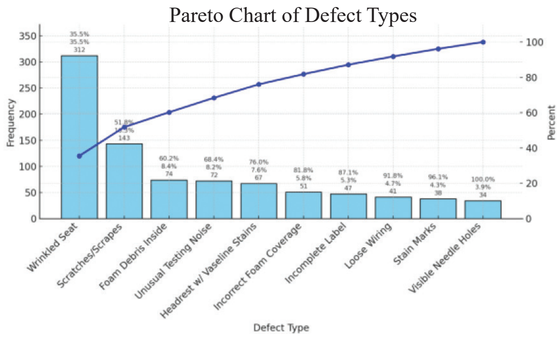

Fig. 2. Pareto Chart of the Top 10 Defect Types in the MMTH 4P00 Production Line (April-June 2023).

C. *Analyze Phase*

Environmental factors, particularly humidity and fabric tension, have been shown to critically influence material behavior during automotive seat production [9].

To better understand the root causes of rear cushion wrinkling in the MMTH 4P00 production line, we applied Lean Six Sigma tools, including the Fishbone Diagram (Cause-and-Effect Diagram) and the 5 Whys method. Data from the Measure phase indicated that "Wrinkled Seat" was the most frequent defect, with 312 recorded cases, accounting for 30.3% of total issues. The Fishbone Diagram (Fig. 3) categorized potential causes into six key areas: Material, Process, Equipment, Environment, People, and Measurement, and a team of production operators, quality control personnel, and process engineers collaboratively examined factors such as material tension inconsistencies, deviations in sewing procedures, and ergonomic difficulties during assembly. By identifying these factors, the team was able to prioritize corrective actions that directly targeted the primary defect. The findings helped develop standardized sewing protocols and improved material handling procedures, ensuring that improvements aligned with Lean Six Sigma principles for defect reduction and process optimization.

The root causes of rear cushion wrinkling were systematically categorized into six key areas:

1) Material: The extruded fleece's excessive stiffness from plastic reinforcement reduced flexibility around curved seat sections, creating stress points.

2) Process: Inconsistent sewing techniques and improper fleece cutting introduced localized tension, worsening wrinkles in high-stress zones.

Equipment: Hog ring tension imbalances and needle misalignment in stitching machines produced uneven seams, amplifying fabric stress.

3) Environment: Humidity fluctuations compromised material elasticity, causing unpredictable contraction/expansion, while poor lighting hindered defect detection.

4) People: Operator handling variability (e.g., inconsistent stretching and pressure application) and

lack of specialized training led to assembly inconsistencies.

5) Measurement: Inadequate quality checks allowed defects to pass undetected, and tolerance deviations in fleece dimensions occasionally exceeded specifications.

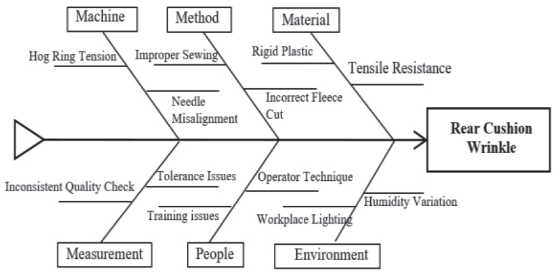

Fig. 3. Fishbone Diagram illustrating potential causes of rear cushion wrinkling in the MMTH 4P00 production line.

To further investigate the root causes of rear cushion wrinkling, the 5 Whys method [1] was applied to the key factors identified in the Fishbone Diagram: material stiffness, improper stitching techniques, and machine-related issues. This iterative approach involved repeatedly asking "Why?" to trace surface-level problems to their underlying causes. The cross-functional team collaborated to validate responses using production data, operator feedback, and machine logs. For instance, questioning why material stiffness occurred revealed inadequate plastic reinforcement specifications, while repeated "Whys" on stitching errors exposed outdated training protocols. The results, summarized in Table II, mapped the causal chain for each factor, providing actionable insights to address systemic gaps in material design, process standardization, and equipment calibration.

TABLE II
WHY-WHY ANALYSIS FOR PRIMARY CAUSES OF REAR CUSHION WRINKLING

Factor	Why 1	Why 2	Why 3	Why 4	Why 5	Root Cause
Material Stiffness	Q: Why does the fleece cause wrinkling? A: The plastic reinforcement makes it too stiff to flex around curved sections.	Q: Why is the fleece too stiff? A: The Plastic reinforcement was selected for durability, not flexibility.	Q: Why was this material chosen? A: It met the cost and longevity requirements during supplier selection.	Q: Why wasn't flexibility prioritized? A: The material specification the process did not account for curvature demands of the MMTH 4P00 design.	Q: Why wasn't curvature considered? A: There was a lack of collaboration between design and procurement teams.	Insufficient cross-functional collaboration during material specification led to the selection of an overly stiff fleece material.
Improper Sewing Techniques	Q: Why do improper sewing techniques cause wrinkling? A: Inconsistent stitching creates stress points on the fleece.	Q: Why is stitching inconsistent? A: Operators use varying techniques for curved sections.	Q: Why do operators use varying techniques? A: There is no standardized sewing procedure for curved areas.	Q: Why isn't there a standardized procedure? A: Process documentation does not address curved stitching challenges.	Q: Why wasn't this addressed in documentation? A: The process was developed without input from experienced operators.	Lack of operator input in process documentation resulted in the absence of a standardized sewing procedure for curved sections.
Machine Issues	Q: Why do hog ring tension and needle misalignment cause wrinkling? A: They Produce uneven stitching, increasing material tension.	Q: Why are hog rings and needles misaligned? A: The machines have not been calibrated regularly.	Q: Why haven't machines been calibrated? A: There is no scheduled maintenance plan forstitching equipment.	Q: Why is there no maintenance plan? A: Maintenance responsibilities were not clearly assigned during production setup.	Q: Why weren't responsibilities assigned? A: Management overlooked the need for a formal maintenance protocol.	The absence of a formal maintenance protocol resulted in uncalibrated machines, which in turn caused inconsistent stitching.

The 5 Whys analysis (Table II) identified three root causes driving rear cushion wrinkling:

1) Material Stiffness: Overly rigid fleece material, selected due to poor collaboration between the design and procurement teams.

2) Process Gaps: Inconsistent sewing techniques caused by incomplete process documentation and training.

3) Equipment Neglect: Uncalibrated stitching machines resulting from the lack of a preventive maintenance protocol.

These findings validated the initial hypothesis of excessive tension in the extruded fleece while exposing systemic flaws in cross-departmental coordination, process standardization, and equipment management. The tabular presentation of the 5 Whys (Table II) clarified causal chains—from surface defects to organizational weaknesses, enabling stakeholders to prioritize targeted interventions. For the Improve phase, solutions will focus on material redesign (e.g., flexible fleece alternatives), standardized sewing workflows, and calibrated maintenance schedules to address root causes holistically.

D. *Improve Phase*

Modifying the listing fleece to improve flexibility aligns with recent research on optimizing seat structure for both comfort and manufacturability [10]. Building on the root causes identified in the Analyze phase—specifically, material stiffness, improper sewing techniques, and machine-related issues—the Improve phase focused on developing and testing solutions to mitigate rear cushion wrinkling in a pickup truck seat production line. The primary intervention targeted the extruded fleece strip, which was found to exhibit excessive tension due to its plastic reinforcement, particularly in curved sections. The team proposed modifying the fleece strip design (length: 370 mm) by cutting rectangular slots to reduce plastic resistance. This adjustment was designed to achieve the project objectives, cutting wrinkle defects by 50% and bringing the defect rate below 2.0%. The redesign prioritized enhancing material flexibility while maintaining structural integrity, ensuring compatibility with standardized sewing processes and calibrated equipment.

1) Cutting the Listing Strip into Rectangular Slots

The intervention began by modifying the listing strip design, incorporating rectangular slots (15×5 mm) to reduce plastic resistance. Two configurations were tested: slots spaced at 60 mm intervals and those at 80 mm intervals. A total of 30 samples—15 for each configuration—were prepared to evaluate their effectiveness in minimizing wrinkling in curved seat sections. Fig. 4 provides a visual comparison of the original and modified listing strips, emphasizing the design adjustments aimed at enhancing material flexibility while maintaining structural integrity. The 60 mm and 80 mm slot spacings were selected based on preliminary design trials and sewing ergonomics. While 60 mm aimed to enhance stress relief, 80 mm offered a better balance between flexibility and material strength. Both values were evaluated under real-world conditions without modifying existing production tooling.

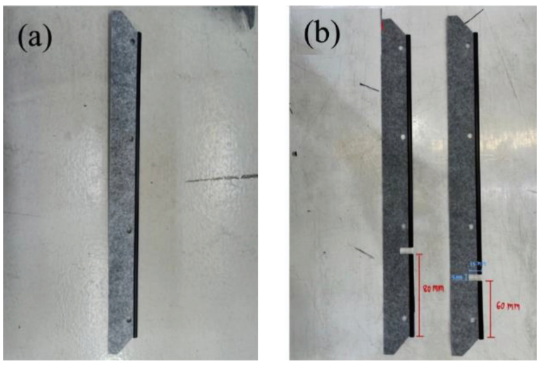

Fig. 4. Comparison of the original listing strip (a) and the modified strips with rectangular slots at 60 mm and 80 mm intervals (b).

2) Sewing Integration

The modified listing strips were integrated into the stitching system, with the 60 mm slots aligned at mark 2 and the 80 mm slots positioned between mark 2 and mark 3—areas susceptible to wrinkling due to material tension. Stitching operators reported that the slotted strips were easier to handle around curves, reducing stitching time. Fig. 5 illustrates this process.

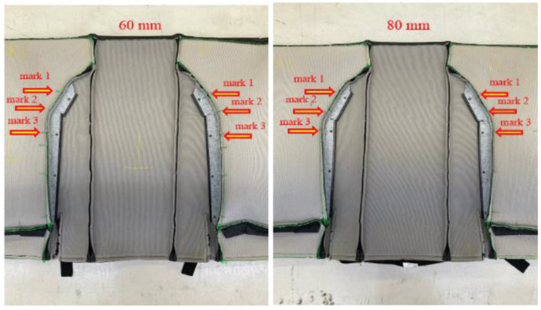

Fig. 5. Sewing integration of the modified listing strips with rectangular slots at 60 mm and 80 mm intervals.

Preliminary observations of the trim's display side showed wrinkling in all configurations (original (a), 60 mm (b), and 80 mm (c)); however, definitive results were deferred until the foam covering stage. Fig. 6 presents these initial observations.

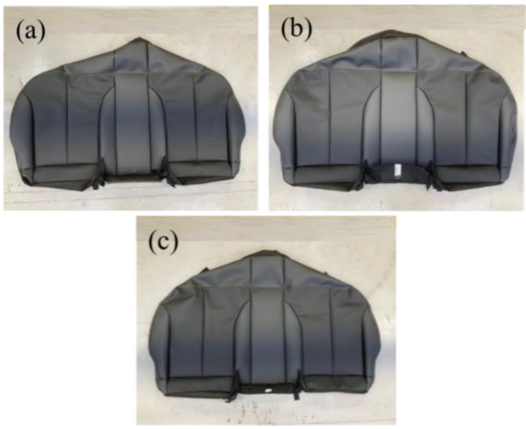

Fig. 6. External appearance of the trim's show side for the original and modified listing strips at 60 mm and 80 mm intervals.

3) Foam Covering

The stitched trims were then attached to the foam cushion to assess the effectiveness of the intervention. Fig. 7 depicts the foam covering process.

Fig. 7. Foam covering process for trims with modified listing strips.

The cushions were evaluated in comparison to the customer's master sample to ensure compliance with quality standards. The original strip exhibited excessive wrinkling, leading to a non-conformance (NG status) due to failure in meeting aesthetic and structural requirements. The 60 mm configuration demonstrated a reduction in wrinkling; however, the defect level remained beyond acceptable limits, resulting in an NG status. Conversely, the 80 mm configuration effectively mitigated wrinkling, achieving compliance with quality standards and receiving an OK status. Fig. 8 presents a comparative analysis of the results.

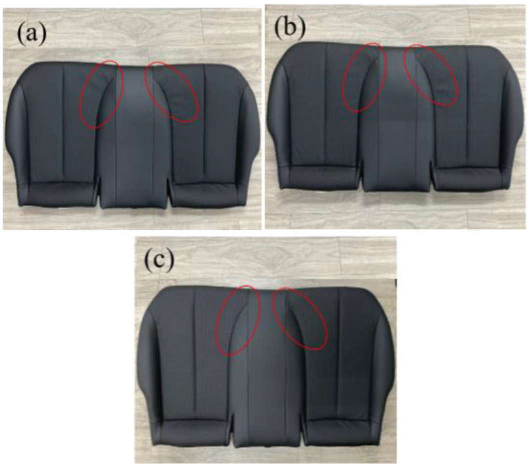

Fig. 8. Comparison of rear cushions after foam covering, using the original listing strip (a) and modified strips at 60 mm (b) and 80 mm (c) intervals.

4) Validation of the 80 mm Configuration

A follow-up trial with 30 additional samples using the 80 mm configuration confirmed its effectiveness, with nearly all samples meeting quality acceptance criteria, aligning with customer specifications, and successfully addressing the wrinkling issue in curved sections. Table III summarizes the initial trial results, showing that all 15 samples with the 60 mm slot spacing failed (100% NG), whereas the 80 mm slot spacing performed significantly better, with 13 out of 15 samples passing (86.67% OK) and only 2 failing (13.33% NG).

TABLE III
INSPECTION RESULTS FOR WRINKLING ACROSS 60 MM AND 80 MM CONFIGURATIONS

Slot Spacing Configuration	OK (Pieces)	NG (Pieces)	Total (Pieces)	OK Percentage	NG Percentage
60 mm	0	15	15	0%	100%
80 mm	13	2	15	86.67%	13.33%

Note: OK, the sample meets the customer's master sample standards with no visible wrinkling; NG: The sample fails to meet standards due to visible wrinkling.

A total of 30 samples were tested, with 15 samples for each slot spacing configuration (60 mm and 80 mm).

Although the difference in performance between the 60 mm and 80 mm configurations was visibly substantial, a two-proportion Z-test was performed to statistically confirm the significance of this improvement. The result ($Z = -4.79$, $p < 0.001$) validated that the 80 mm slot spacing significantly improved the proportion of acceptable parts, supporting the selection of this design for implementation.

E. *Control Phase*

After successfully implementing the 80 mm slotted listing strip in the Improve phase, the Control phase was implemented to sustain improvements and to prevent rear cushion wrinkling from recurring. Several control measures were implemented to maintain reduced defect rates and cost savings.

1) Standardization of Processes

The 80 mm slotted listing strip design was standardized for the pickup truck seat production line. Process documentation was updated to detail the cutting and sewing procedures for the modified strip, including precise slot measurements (15 × 5 mm) and alignment points (between mark 2 and mark 3). Sewing operators were trained in the updated procedure to ensure consistency, mitigating the improper sewing techniques identified in the Analyze phase. An SOP manual was distributed to all relevant personnel, and regular audits were scheduled to ensure adherence to the new standard.

2) Equipment Maintenance Protocol

To mitigate machine-related issues, a structured maintenance protocol was implemented for stitching machines. A scheduled monthly calibration plan was introduced, with designated maintenance responsibilities assigned to a specialized technician team. Calibration checklists were introduced to monitor maintenance schedules and ensure hog rings and needles met specified tolerances, preventing uneven stitching, which could lead to wrinkling defects.

3) Monitoring and Control Chart

A control chart was introduced to systematically monitor defect rates after the intervention. Daily defect counts for wrinkling were recorded and plotted against established control limits, with a target defect rate below 2.0%. This enabled the team to promptly identify deviations and implement corrective measures, such as re-inspecting listing strips or recalibrating machinery.

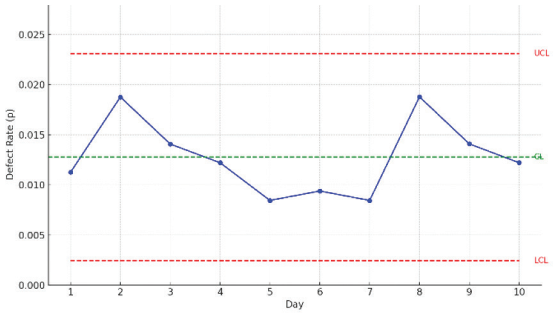

Fig.9. p-Chart for Wrinkling Defect Rate (October 2023, Sample of 10 Days)

This control chart illustrates the daily proportion of wrinkling defects after implementing the 80 mm slot design. The defect rates remained within the control limits (UCL and LCL), indicating that the process was statistically stable during the post-improvement period. Quality control personnel underwent training to effectively utilize the control chart, ensuring continuous monitoring and process stability.

V. Results and Discussion

A. Results

Successful Lean Six Sigma projects require robust change management and training strategies, especially when introducing new work processes [11].

The Lean Six Sigma project targeting rear cushion wrinkling in the MMTH 4P00 production line at a sample company in the automotive manufacturing industry led to notable enhancements in both quality and financial performance, as evaluated using the DMAIC framework. The following sections outline the project's outcomes concerning defect reduction and cost savings, validated through production data and quality inspection reports

Baseline defect rate was established at 4.05% during the Define and Measure phases, with 1,028 defective units identified out of 25,408 produced between April and June 2023, averaging 342.7 defective units per month. The primary cause was excessive tension in the extruded listing fleece, particularly in curved seat sections, leading to rear cushion wrinkling. Through root cause analysis and targeted process improvements—specifically, modifying the listing fleece design with rectangular slots (15 × 5 mm) at 80 mm intervals—the defect rate was lowered to 1.61% by October 2023. This marks a 60.27% reduction in the defect rate (from 4.05% to 1.61%), exceeding the project objective of achieving a rate below 2.0%. Specifically for wrinkling defects, the count decreased by 40%, from 312 pieces to 187 pieces over the intervention period.

Table IV compares the monthly defect counts before (April-June 2023) and after (July-October 2023) the intervention. Before the intervention, the average monthly defect count stood at 342.7 pieces. After the intervention, this number dropped to 171.5 pieces, reflecting a 50% reduction in overall defects. Statistical Process Control (SPC) tools, including control charts and Pareto analysis, confirmed process stability, with the sigma level improving from 3.2 (66,807 DPMO) to 4.1 (6,210 DPMO) according to standard conversion values [1], indicating enhanced process capability. Furthermore, cycle time was reduced by approximately 20%, enhancing overall production efficiency.

TABLE IV
DEFECT COUNTS IN PRODUCTION LINE BEFORE AND AFTER IMPROVEMENT

Period	Month	Total Produced (Pieces)	Defect Count (Pieces)	Defect Rate (%)
Before Improvement	April	8,469	319	3.77%
	May	8,469	380	4.49%
	June	8,470	329	3.88%
	April-June	25,408	1,028	4.05%
After Improvement	July	10,652	204	1.92%
	August	10,652	189	1.77%
	September	10,652	157	1.47%
	October	10,653	136	1.28%
	July - October	42,609	686	1.61%

Reducing the number of defects in the manufacturing process directly results in cost reduction, with the unit cost of the rear seat production being 6,820 baht (During April to June 2023). There were 312 cushions with wrinkles found, representing a cost of 2,127,840 baht (312 × 6,820 baht). After improving (During July to October 2023), the number of cushions with wrinkling defects decreased to 187, reducing costs to 1,275,340 baht (187 × 6,820 baht). The cost savings of 852,500 baht, or 54.81%, exceeded the original target of 40%. This was primarily achieved through reduced rework and more efficient material usage. In addition, the reduced cycle time made the production process more agile and flexible. Additionally, customer satisfaction increased approximately 20% according to Voice of Customer (VOC) surveys, reflecting improved product quality. And consistency with Critical to Quality (CTQ) requirements.

B. Discussion

Successful Lean Six Sigma projects require robust change management and training strategies, especially when introducing new work processes [11]. This Lean Six Sigma project demonstrated the effectiveness of integrating Lean's waste-reduction principles with Six Sigma's data-driven methodology in addressing defects in automotive seat production. Root cause analysis was conducted using tools such as the Fishbone Diagram and Why-Why Analysis, which systematically identified key contributing factors. To mitigate this issue, the listing fleece design was modified by incorporating rectangular slots (15 × 5 mm), spaced at 80 mm intervals. This modification yielded a 60% reduction in wrinkling defects, decreasing from 312 to 187 pieces.

As of October 2023, the overall defect rate has decreased to 1.61%, exceeding the target of 2.0%. This result is consistent with past research, which has shown that Lean Six Sigma methods can reduce defect rates in automotive part manufacturing by 50-70% [2]. The results of this research study truly demonstrate the effectiveness of a Lean Six Sigma problem-solving approach.

However, some important issues should be further considered in future research, especially during the improvement process. The sewing workers had to adapt to the new process at first, which felt unfamiliar and uncomfortable. Therefore, to solve this problem, additional training, including practical training, was provided to make the workers more confident and able to follow the new steps correctly and efficiently. This data reflects the importance of effective change management in Lean Six Sigma projects [12].

Another challenge was the lack of real-time defect tracking, which resulted in delays in identifying issues. Additionally, limited data on cycle time and operator efficiency restricts in-depth analysis. These gaps highlighted the need for future improvements, such as implementing automated systems to monitor cycle time and operator performance in real time.

During the Control phase, the defect rate was maintained at 1.61% through regular audits and SPC, ensuring process stability. However, environmental factors such as humidity, identified through the Fishbone analysis, could pose future challenges, necessitating continuous monitoring and adjustments as needed.

VI. Conclusion

A. Summary

This Lean Six Sigma project addressed rear cushion wrinkling in a production line at a sample company in the automotive industry, demonstrating significant improvements using the DMAIC methodology. Wrinkling defects were reduced by 60%, from 312 to 187 pieces, lowering the overall defect rate from 4.05% to 1.61% within six months, surpassing the target of reducing it below 2.0%.

Identifying excessive tension in the listing fleece as the primary issue was crucial. Several modifications were implemented, including adding rectangular slots (15 × 5 mm, spaced 80 mm apart), standardizing sewing procedures, and incorporating SPC tools to ensure process stability.

The cost savings from wrinkling totaled 852,500 THB, which represented a 54.81% reduction, surpassing the original target of 40%. The process resulted in a 20% cycle time reduction while customer satisfaction increased by 20%, maintaining adherence to Critical to Quality (CTQ) like dimensional accuracy. The study by [2] presents clear evidence that Lean Six Sigma methods can lead to substantial improvements in quality, efficiency, and profitability within automotive seat production facilities.

B. Recommendations

To ensure the sustainability of these improvements, the following best practices should be adopted:

Continuous Training Programs: Conduct structured and ongoing training for operators on updated work standards and quality control procedures to ensure process consistency and minimize variability, mitigating initial operator resistance observed in the Improve phase.

• Real-Time Monitoring and Audits: Establish a structured system for regular quality audits and real-time defect tracking using SPC tools, such as control charts, to detect deviations early and maintain long-term stability, a proven practice in Lean Six Sigma implementations [1].

• Culture of Continuous Improvement: Embed a culture of continuous improvement within the organization to reduce resistance to change and support adherence to Lean Six Sigma principles, promoting long-term commitment from all stakeholders.

Future research could focus on predictive analytics to predict defect trends by leveraging historical control chart data to forecast wrinkling occurrences. Additionally, future studies could assess the scalability of these findings to other automotive components, such as front seat assemblies, to extend the applicability of Lean Six Sigma across various production processes.

C. Future Improvements

Future work could integrate AI-based real-time defect monitoring to further enhance quality control, consistent with current trends in industrial process monitoring [13]. The success of this project paves the way for expanding Lean Six Sigma methodologies to additional defect categories in the 4P00 production line, such as stitching inconsistencies, material delamination, and assembly misalignments. Utilizing the DMAIC framework, beginning with root cause analysis using tools such as the Fishbone Diagram, could achieve similar defect reductions and cost savings while enhancing overall product quality.

Furthermore, automation presents a significant opportunity for defect detection. IoT sensors combined with machine learning algorithms enable real-time tracking of fabric tension and sewing precision while monitoring environmental humidity, which was determined as a key factor in Fishbone analysis [14]. IoT sensors enable identification of excessive tension in listing fleece during production, which allows for real-time adjustments to prevent wrinkling.

Pilot studies could evaluate automated vision systems for real-time detection of wrinkles and other defects, with the potential to elevate the sigma level beyond 4.5 (e.g., reducing DPMO below 1,350), thereby establishing a new quality benchmark in automotive manufacturing.

Acknowledgment

We extend our sincere gratitude to the production team at a sample company in the automotive industry for their invaluable support, particularly for granting us access to their facility. Special appreciation is extended to the sewing team leader and the sewing team for their assistance in testing and for providing valuable feedback, as well as the QC team for ensuring the accuracy and completeness of our defect data.

We are also profoundly grateful to our university advisor for their invaluable guidance, which significantly contributed to the successful completion of this research.

References

[1] T. Pyzdek and P. Keller, *The Six Sigma Handbook*, 4th ed. New York: McGraw-Hill Education, 2014, pp. 1-814.

[2] J. Antony, S. Gupta, and V. Sunder, *Lean six sigma in manufacturing: A review*, London, UK: Intech Open, 2020, pp. 1-29.

[3] M. L. George, *Lean six sigma: Combining six sigma quality with lean speed*. New York: McGraw-Hill, 2002, pp. 1-830.

[4] J. C. Tsou and C. W. Chen, "Applying six sigma methodology to optimize the manufacturing process of car seats: A case study on defect reduction," *Int. J. Ind. Eng. Theory Appl. Pract.*, vol. 12, no. 3, pp. 245-254, 2005.

[5] H. H. Purba and A. Sunadi, "Quality improvement in the car seat industry using Quality Function Deployment (QFD)," *J. Ind. Eng. Manag.,* vol. 11, no. 4, pp. 672-689, 2018.

[6] S. Gupta and R. K. Jain, "Application of Statistical Process Control (SPC) in automotive manufacturing: A case study on defect reduction," *Int. J. Quality Rel. Manag.*, vol. 33, no. 5, pp. 678-695, 2016.

[7] D. C. Montgomery, *Introduction to statistical quality control,* 8th ed. New Jersey: Wiley, 2020, pp. 1-729.

[8] Yamada Machine Tools, "Eliminate defects in processing unmanned production line for complex-shaped parts," *SMRI Asia*, 2021. [Online]. Available: https://www.smri.asia/th/yamadamtt/news/4837 [Accessed: Mar. 15, 2024].

[9] H. Park, S. Lee, and J. Kim, "Impact of environmental conditions on fabric tension in automotive seat manufacturing," *J. Ind. Text*., vol. 47, no. 7, pp. 1198-1215, 2018.

[10] K. Chung and Y. Kim, "Design optimization of automotive seat structure for occupant comfort and manufacturing efficiency," *Int. J. Automot. Technol*., vol. 20, no. 3, pp. 561-569, 2019.

[11] M. Lande, R. L. Shrivastava, and D. Seth, "Critical success factors for lean six sigma in SMEs," *J. Manuf. Technol. Manag.*, vol. 27, no. 4, pp. 446-467, 2016.

[12] J. P. Kotter, *Leading Change*. Boston, MA, USA: Harvard Business Review Press, 2012, pp. 1-10.

[13] S. Yin, S. X. Ding, A. Haghani, H. Hao, and P. Zhang, "A review on basic data-driven approaches for industrial process monitoring," *IEEE Trans. Ind. Electron*., vol. 61, no. 11, pp. 6418-6428, 2014.

[14] Z. Li, Y. Zhang, and L. Wang, "Real-time defect detection in manufacturing using IoT and machine learning: A case study in automotive production," *IEEE Trans. Ind. Inform*., vol. 17, no. 8, pp. 5432-5441, 2021.

Bundit Wongthong is a Lecturer in the Department of Industrial Engineering at Thonburi University, Bangkok, Thailand. His research interests include Lean Six Sigma methodologies and process optimization in manufacturing systems.

Poom Jatunitanon is a Lecturer in the Automotive Manufacturing Engineering Program at Panyapiwat Institute of Management, Bangkok, Thailand. He focuses on automatic control systems, PLC control, vibration, and damage analysis of automotive parts.

Bundit Inseemeesak is an Assistant Professor in the Automotive Manufacturing Engineering Program at Panyapiwat Institute of Management, Bangkok, Thailand. His research interests lie in biomass processing, the production of natural composite materials, and engineering design.

Yodnapha Ketmuang, the Corresponding Author, is a Lecturer in the Automotive Manufacturing Engineering Program at Panyapiwat Institute of Management, Bangkok, Thailand. Her research focuses on Statistical Process Control (SPC) and Lean Manufacturing applications in the automotive industry.

Enhancing Thai Rice Query Assistance through a Knowledge-Driven Approach Using GraphRAG

Gampanut Soontontam[1], Tinnaphob Dindam[2], Adisorn Kheaksong[3], Kanabadee Srisomboon[4], and Parinya Sanguansat[5*]

[1,2,3,5]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand

[4]Department of Electrical and Computer Engineering, King Mongkut's University of Technology North Bangkok, Bangkok, Thailand

E-mail: 6652300168@stu.pim.ac.th, tinnaphobdin@pim.ac.th, adisornkhe@pim.ac.th, kanabadee.s@eng.kmutnb.ac.th, parinyasan@pim.ac.th*

Received: February 28, 2025 / Revised: April 4, 2025 / Accepted: May 19, 2025

***Abstract*—Thai rice farmers face significant challenges accessing timely and accurate information for crucial decisions regarding variety selection, soil management, and adapting to climate change. While Retrieval-Augmented Generation (RAG) systems aim to provide information, traditional RAG often struggles with complex queries requiring interconnected knowledge and can yield generic or less relevant answers in specialized domains like agriculture due to its reliance on the semantic similarity of isolated text chunks. This paper introduces and evaluates GraphRAG, a knowledge graph-enhanced RAG approach, designed specifically to overcome these limitations and improve query assistance for Thai rice cultivation. The methodology involves constructing a knowledge graph from key Thai rice farming documents and integrating it with a large language model to provide context-aware responses, comparing its performance against a traditional RAG baseline. Results demonstrate GraphRAG's superior effectiveness; user preference tests showed participants favored GraphRAG responses (52.9%) significantly more than traditional RAG (35.3%), particularly for complex queries requiring nuanced understanding. Quantitatively, GraphRAG showcased enhanced efficiency, reducing the average query response time by nearly 3 times (from 1.43 seconds for RAG to 0.41 seconds) and decreasing memory usage by over 50% (from 457.42 KB for RAG to 213.09 KB). This study concludes that GraphRAG offers a valuable approach for enhancing information retrieval accuracy, contextual understanding, and system efficiency in specialized, low-resource agricultural domains, highlighting its significance for providing better decision support to farmers.**

***Index Terms*—GraphRAG, LLM, Thai Rice, Knowledge Graph, Retrieval-Augmented Generation**

I. Introduction

Rice (Oryza sativa L.) is the main food for more than half of the world's population. Mostly grown in tropical zones, rice is full of carbohydrates, proteins, and healthy compounds, which greatly help with various health benefits. Recent improvements in rice science, especially in Androgenesis, have made breeding methods better, but issues like dependence on the type of plant still exist. Besides its role in food security, rice's nutrition, which includes important vitamins and minerals, makes it a key part of many food items and health supplements [1]. Also, farming actions, like when planting is done, greatly affect how much rice is produced, highlighting how important it is to make growing methods as good as possible [2]. Using technology, such as machine learning to identify seed types, in rice farming is also causing big changes [3]. These points together show how rice is important in many ways for both farming and public health, which will be talked about more in the next parts. Growing rice is very important in Thailand, as both a major economic product and a cultural symbol. Thailand is one of the top countries selling rice worldwide, with types like Jasmine rice (Oryza sativa var. Indica) becoming known around the world for their special smell. However, rice farmers have many problems, including picking the best rice types, managing their soil, and adapting to climate change. Normal ways of sharing information, like agricultural extension services, are often slow and not very effective, which stops farmers from getting timely and useful information [4]. To fix these problems, ways that use knowledge and rely on Artificial Intelligence (AI) and Machine Learning (ML) are becoming more and more needed in farming [5].

The rise of Large Language Models (LLMs) has played a very important role in natural language processing, demonstrating remarkable capabilities in generating coherent text, answering complex questions, and even producing creative content. Their ability to learn from massive datasets has positioned them as powerful tools across diverse applications. However, despite their impressive general knowledge, LLMs inherently suffer from a critical limitation: a lack of up-to-date, specialized, and contextually relevant domain knowledge [6]. This deficiency becomes particularly pronounced when addressing tasks requiring precision and deep understanding within specific fields, especially in agriculture.

Traditional methods of fine-tuning LLMs on domain-specific datasets can be computationally expensive. Practically, Retrieval-Augmented Generation (RAG) systems offer good answers by mixing the ability to find documents with the language power of Large Language Models (LLMs) to create correct and relevant responses by retrieving relevant documents or information snippets at runtime [7]. However, normal RAG systems often have trouble with questions about specific topics because they are not very good at understanding complex relations between pieces of data and entities [8]. To overcome this problem, systems like GraphRAG use knowledge graphs to produce a graph-structured knowledge representation that depicts domain hierarchies and entity interactions [9]. The design of these knowledge-based solutions will be looked at closely throughout this paper.

The GraphRAG system is designed to solve these issues by including a knowledge graph. This organizes data into connected items and links, thus helping to give more thoughtful and helpful answers to questions [10]. In this paper, the Llama 3 8b model is utilized as the Large Language Model (LLM), and Neo4j is used as a graph representation, which has shown it is good at understanding and creating natural language. GraphRAG can make more precise and contextually suitable answers. Furthermore, we use LangChain to help the system's ability to take out useful data from PDF documents, adding to the knowledge graph and allowing for more complete handling of questions [11]. This paper introduces GraphRAG, a knowledge graph-based extension of Retrieval-Augmented Generation (RAG), designed to enhance the retrieval and generation of Thai rice-related information.

We compare GraphRAG with traditional RAG models in this specific task. The remainder of this paper is organized as follows: Section II reviews related work in knowledge graphs and RAG systems, Section III details the proposed GraphRAG methodology and experimental setup, Section IV presents and discusses the results of our comparative analysis, and Section V concludes the paper with key findings and future research directions. The results demonstrate that GraphRAG not only outperforms standard RAG systems, particularly when handling complex queries and extracting useful insights from unstructured data sources, but also remains effective even in low-resource settings within the Thai rice domain.

II. Related Work

Past research has demonstrated the value of incorporating knowledge graphs (KGs) to enhance information searching and Question-Answering (QA) systems across various domains. However, traditional QA systems and even standard Retrieval-Augmented Generation (RAG), which retrieves text chunks based on vector similarity to augment Large Language Models (LLMs), can face limitations in handling complex queries requiring deep contextual understanding or reasoning over interconnected information [8].

Recent advancements have led to the development of GraphRAG, an approach that specifically integrates KG construction and querying within the RAG framework to address these limitations [9]. GraphRAG has been explored in various contexts, such as improving accuracy in complex domains like healthcare support [12] or enhancing document analysis through structured knowledge extraction [13]. Unlike standard RAG, GraphRAG typically involves dynamically extracting entities and relationships from the source documents themselves to build a knowledge graph, which then provides richer, interconnected context for the LLM's response generation.

While GraphRAG offers a novel integration, other research has focused on leveraging pre-existing or separately constructed KGs. For example, the UCKG-Why-QA system demonstrated the utility of a cause-and-effect KG for diagnosing plant diseases, including rice sicknesses, showing improved accuracy for complex questions through visualization [15]. However, it encountered challenges with fine-grained farming details. Similarly, Xie *et al.* [16] utilized Neo4j and BERT-CRF to build a dedicated TCM knowledge graph (with over 2,200 entities and 5,100 relationships), enhancing specialized searches but facing limitations due to data availability.

Comparing these approaches, GraphRAG differs significantly. While UCKG-Why-QA and the TCM system rely on potentially pre-defined or separately curated KGs, GraphRAG often constructs its knowledge graph directly from the input text corpus used for retrieval. This tight integration aims to provide context that is highly relevant to the source documents and potentially offers greater scalability and adaptability compared to methods requiring extensive manual ontology engineering or facing the data completeness issues noted by Xie *et al*. [16]. Furthermore, GraphRAG's emphasis on relationship traversal within the generated graph provides a mechanism for deeper reasoning compared to the semantic chunk retrieval of standard RAG. While other methods also exist to enhance RAG, such as sophisticated re-ranking algorithms applied after initial retrieval [14], this study focuses specifically on the potential of the integrated knowledge graph construction offered by GraphRAG.

Therefore, this research investigates the application and effectiveness of the GraphRAG approach within the specific, low-resource domain of Thai rice farming, comparing it against a standard RAG baseline to evaluate its benefits in addressing complex farmer queries.

III. Methodology

A. Knowledge Graph Visualization

To illustrate how different parts of the Thai rice system are linked and function together, we have constructed a knowledge graph. This graph is essentially a visual representation that utilizes nodes and relationships to organize information. In this paper, we built the knowledge graph as shown in Fig.1 from our data collection, which will be explained later. However, it is hard to view in detail because of its size. We show some of the chunks as follows:

Fig. 2 is the Thai rice industry knowledge graph; the nodes and edges in this graph are:

• Business People: The main part is the Thai Rice Business; the lines connect it to important groups that research Thai Farmers and groups that study rice selling.

• Countries: Thailand is shown along with countries that are its neighbors and also sell rice (Cambodia, Myanmar, Vietnam, India, and Pakistan). This shows who competes with whom and who trades with whom.

Types of Connections: The lines (with labels) show different kinds of connections. Here's an example of the connections:

• EXPORTS TO: This Shows where Thai rice is exported.

• COMPETITO: Points out which countries are competitors with Thailand in selling rice.

• COMPARED TO: Shows when Thai rice is being compared to or studied against something.

This graph helps us understand better how the Thai rice business is set up, how rice flows in trade, and who the competitors are.

Fig. 3 is the Thailand Agricultural Knowledge Graph. The nodes and edges in this graph are:

• GOVERNMENT AND TEAMS: The main part is the Agriculture Ministry, and the lines connect it to important teams like the Agriculture Department, Trade Department, and the Bank for Agriculture. This shows the government teams that help and manage Thai farming.

• MAIN FARM PRODUCT & NATION: Thai Rice is the main product; it's connected to Thailand, showing rice is grown there. Thailand is in the middle, linked to all parts of its farming, showing it runs and helps farming overall.

Types of Connections: The lines (with labels) show different connections. Here are some examples:

• HAS BANK: Shows the Agriculture Ministry links to the Bank for Agriculture, like it's in charge or related.

• HAS AGENCY: Shows the Agriculture Ministry is in charge of Thai Rice.

• PRODUCES: Show what's made in Thailand, like Thai Rice is made in Thailand, and Farming makes food in Thailand.

This graph helps us see how Thai farming is set up, showing how government teams, offices, banks, and farm products are all linked in the country.

Fig. 4 is the Rice Processing Knowledge Graph; the nodes and edges in this graph are:

• RICE COMPONENTS: Milled Rice is the main focus, with related parts around it, such as Rice Husk, Rice Bran, Sticky Rice, and White Rice Seed. These show the different types of rice or materials used when processing rice. This describes the various rice materials involved.

• STEPS IN RICE MAKING: You can see Processing and Contamination. Processing is the way rice is changed from its seed to what we eat. Contamination refers to when rice becomes unclean or contains unwanted substances. These points out the important actions and possible problems during rice production.

Types of Connections: The lines with labels show different kinds of relationships:

• CONTAMINATION: Indicates Milled Rice and Sticky Rice can experience Contamination, meaning they can become impure.

• PROCESSING: This shows that Milled Rice is made through Processing, which changes it to White Rice Seed.

• CONTAINS: Suggests Professor CONTAINS information about Milled Rice, implying experts milled rice.

This graph gives a structured overview of rice processing, covering the rice materials, the steps involved, the knowledge sources, and potential quality issues, all in simple terms.

Fig. 5 is the Rice Disease Knowledge Graph; the nodes and edges in this graph are:

• RICE DISEASES: The main point is that Rice Diseases, Bakanae, Fusarium Moniliforme, and "Virus" are neighbors. These are examples of different sicknesses that can affect rice plants. This shows the types of diseases rice can get.

• CAUSES OF DISEASE: Fusarium Moniliforme and Virus. These are things that CAUSE or CONTRIBUTE TO the Rice Disease happening in rice plants. This tells us what makes rice become diseased.

• SPREADING DISEASE: There is an Insect Vector. This shows how a Virus can be transmitted by insects. Insects can move the "Virus" around to different rice plants. This shows how some rice diseases can spread.

Types of Connections: The lines with words explain the links between them:

• CAUSED BY: Shows that Rice Disease is CAUSED BY Fusarium Moniliforme—this fungus makes the disease happen.

• CAUSES: Shows that Fusarium Moniliforme CAUSES Bakanae—this fungus leads to another disease called Bakanae.

• CONTRIBUTES TO: Shows that Virus

• CONTRIBUTES TO Rice Disease—Viruses are also part of the rice disease problem.

This graph helps to understand rice diseases, what causes them, how they spread, and the different kinds that can affect rice crops.

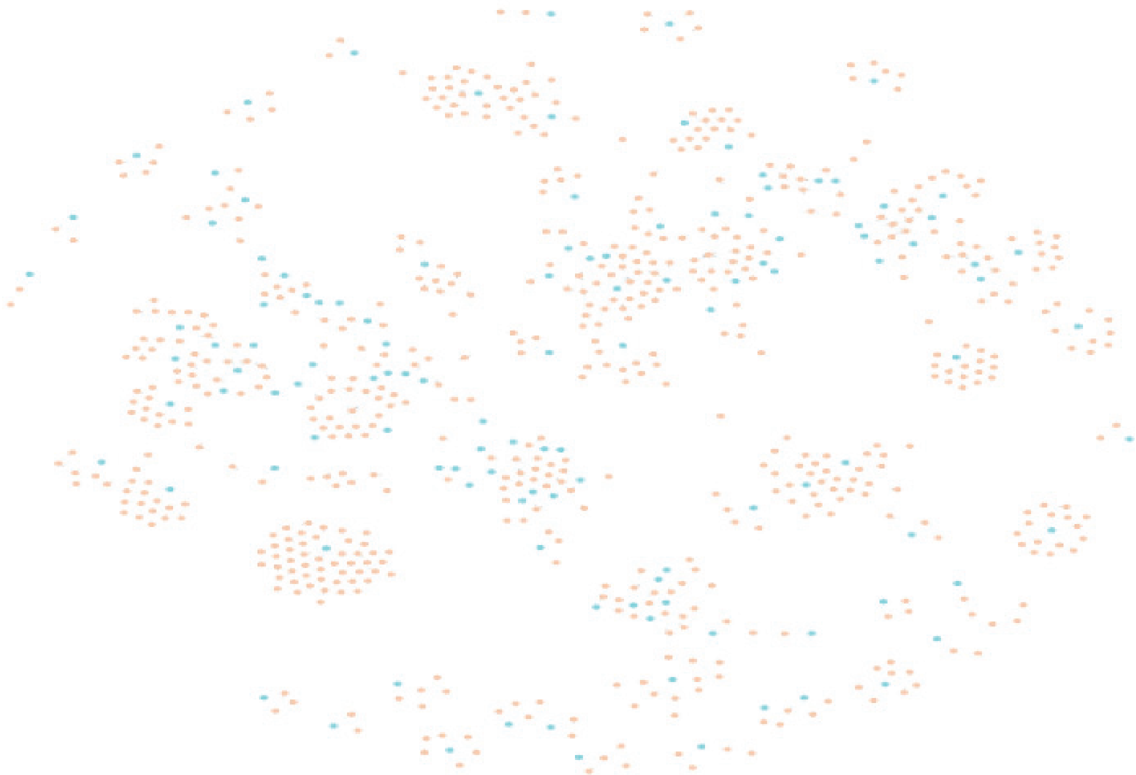

Fig. 1. The knowledge graph for Thai rice that includes hundreds of chunks with nodes and edges visualization using neo4j graph database

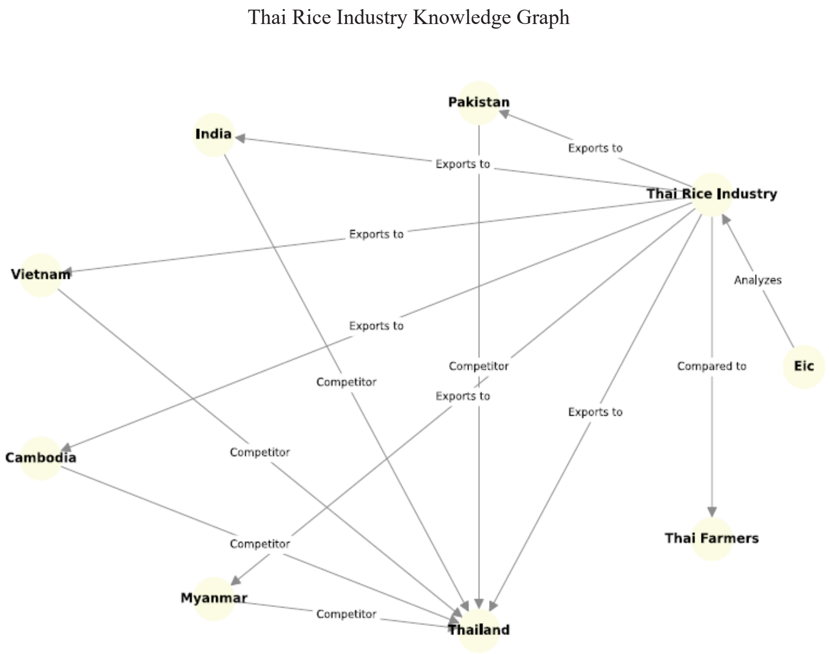

Fig. 2. Thai rice industry knowledge graph. It shows the important people involved, who sells rice to whom, how they compete, and compares them with other countries in Southeast Asia. Dots in the picture are things like Thailand, Thai Farmers, and Vietnam. Lines show connections like EXPORTS TO, COMPETITOR, COMPARED TO.

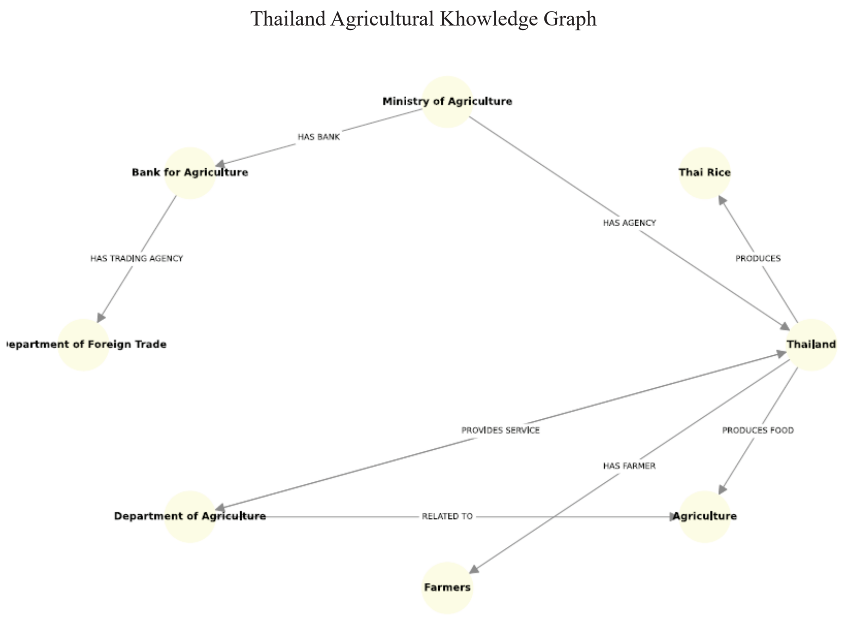

Fig. 3. Thailand Agricultural Knowledge Graph. It depicts key entities within the Thai agricultural domain, such as government ministries, agricultural sectors (like Thai Rice), and stakeholders (e.g., Farmers). Relationships (e.g., HAS BANK, HAS AGENCY, PRODUCES) between these entities are also visualized.

Rice Processing Knowledge Graph

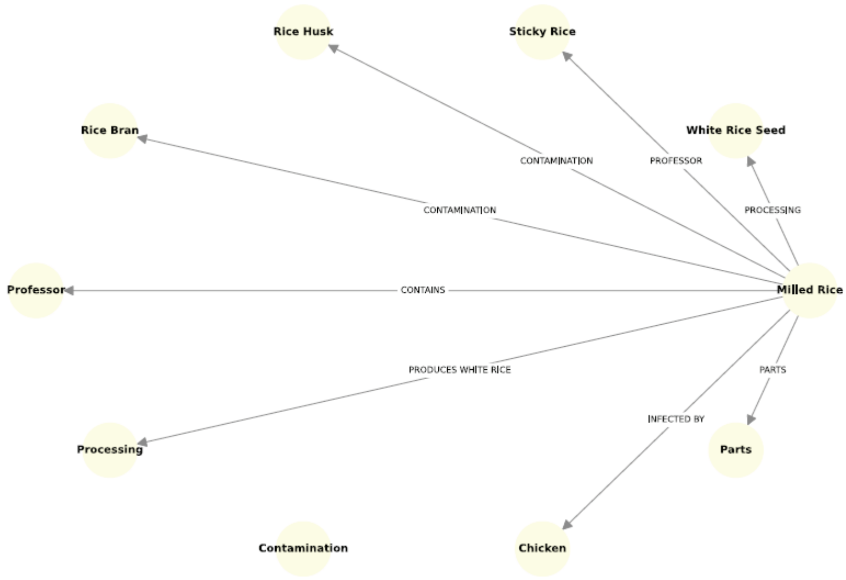

Fig. 4. Rice Processing Knowledge Graph. It visualizes the knowledge graph focused on rice processing stages and related entities. Entities include Rice Husk, Milled Rice, Sticky Rice, Rice Bran, and White Rice Seed, as well as concepts like Processing and Contamination. The Relationships shown are CONTAINS, PRODUCES WHITE RICE, INFECTED BY, and CONTAMINATION.

B. How the Knowledge Graph is Set Up

The GraphRAG system uses a knowledge graph, which we can think of as $G = (V, E)$. Here, V is like a list of nodes (or things), such as different kinds of rice, types of soil, and ways of growing rice. And E is like a list of edges (or relations) that show how these things are connected. These lines have labels to say what kind of connection it is. For example, a line $e = (v_i, \text{PART\_OF}, v_j)$ could mean that a type of rice v_i is part of a certain group or has a special feature linked to v_j [17].

This setup lets GraphRAG use Neo4j's labeled lines to create and follow connections easily. This helps it understand complex information about Thai rice. We can also think of the knowledge graph using a table called an adjacency matrix, where each element is:

$$A_{ij} = \begin{cases} 1 & \text{if there is the line from thing } v_i \text{ to thing } v_j \\ 0 & \text{otherwise} \end{cases}$$

This helps GraphRAG find connected pieces of information that are needed to answer questions, especially when questions are about specific topics where connections are very important for giving good answers [17].

C. Using Cipher Language for Math

To make GraphRAG better at understanding connections in the knowledge graph, we use Neo4j's Cypher language to do some math on the graph. Cypher lets GraphRAG find complex connections, gather information together, and do calculations on the dots and lines in the graph [14].

For example, to find the **total strength** of connections between linked dots (showing how strong or common connections are), we use this Cypher command:

```
MATCH (n) - [r: RELATIONSHIP_TYPE] -> (m)
RETURN n.name, SUM (r.weight) AS totalWeight
```

This command adds up the strengths of all connections between dots (n) and (m) for a certain type of connection. This helps GraphRAG measure and understand how important these connections are when it makes answers [17].

Also, to find the average distance between dots in the knowledge graph (which shows how close things are, like how many steps between a rice type and its features), we use this Cypher command:

```
MATCH p = (a)-[*] -> (b)
RETURN a.name, b.name, AVG (length (p))
AS avgPathLength
```

This command calculates the average distance across different paths between dots. This helps GraphRAG figure out how close and relevant things are when it is answering complex questions that need context [17].

The way we did this study is focused on making and testing the GraphRAG system. GraphRAG is a tool that uses knowledge to help answer questions about Thai rice using a method called Retrieval-Augmented Generation (RAG). The system uses several advanced tools, including Neo4j to manage the knowledge graph, LangChain to get text from PDF files, and the Llama 3 8b Large Language Model

(LLM) to create natural-sounding language. The parts below explain the main pieces of this system and how we tested how well it works.

D. Getting the Data

To create a strong information base, GraphRAG used data from trusted sources for Thai farming information. We got important details from PDF files. These files gave us the key facts needed to answer questions about Thai rice:

• Thai Rice Department: Gave basic info about rice types, how to grow rice, and rules from the government that affect the Thai rice business.

• Thai Agriculture Department: Gave data on taking care of soil, controlling bugs and plant diseases, and good farming methods that help the environment, which are all important for growing rice.

• Kasetsart University, Agriculture Faculty: Shared research from scientists on making better types of rice, improving rice plants through genetics, and getting the best rice harvests.

• Rice Disease Info Files: Gave specific help on finding, handling, and stopping rice diseases.

• Other PDF Files: Had info on rules for selling rice to other countries, how the rice market is changing, and how to deal with climate change in Thai rice farming.

We used LangChain to take the text out of these files and then organized it into a knowledge graph using Neo4j. This helps GraphRAG give clear and correct answers about farming methods, rice plant health, and money matters in Thai rice growing.

E. Examples for Testing

Besides the numbers, we also asked people what they thought. People liked GraphRAG because it gave short, focused answers, especially for hard questions that needed to know how different things are connected. However, for some simpler questions where people expected more explanation, they liked the normal RAG systems better because they gave more details. Knowing this helps us make GraphRAG better in the future. We need to find the right balance between giving quick answers and giving enough information.

Example 1

Question: พื้นที่ปลูกข้าวที่ใหญ่ที่สุดอยู่ที่ไหน? (Where is the largest rice-growing area?)

RAG Response: Thai rice is grown throughout the country, but the largest rice-growing areas are in the Central and Northeastern regions.

GraphRAG Response: India.

Feedback: GraphRAG is better – The user preferred the concise answer from GraphRAG, despite it being factually incorrect. However, the user's preference was likely due to the perceived directness of the response.

Example 2

Question: แนวโน้มการผลิตข้าวในประเทศไทย? (Rice production trends in Thailand?)

RAG Response: Between 1985 and 2020, rice production per hectare of Thai farmers increased by an average of 0.8.

GraphRAG Response: Thailand has a tendency to increase rice production in 2022 with a rice production volume of approximately 15.5 million tons.

Feedback: RAG is better – The RAG system provided a more detailed and contextually appropriate definition, whereas GraphRAG's response was too brief and vague.

Example 3

Question: ข้าวมีกี่สายพันธุ์ ต่างกันอย่างไร? (How many varieties of rice are there? How are they different?)

RAG Response: There are about 40-50 types of Thai rice. Each type of rice has different colors, flavors, and characteristics.

GraphRAG Response: Rice has many species, divided into 2 large groups: Oryza sativa group, which is rice used for direct consumption, and Oryza glaberrima group, which is rice used for making glutinous rice. Rice used for direct consumption is divided into 3 species: Sticky rice, which is sticky and thick; Aromatic rice, which is fragrant and has a sweet taste. Regular rice is a mixture of Sticky rice and Aromatic rice. Each species has different characteristics and properties. Glutinous rice is used to make Sticky rice. It is useful for making Sticky rice.

Feedback: GraphRAG is better—The GraphRAG System provides more descriptive and specific species names than the RAG, which provides general information.

Example 4

Question: ข้าวหอมมะลิแดงที่ปลูกในภาคใต้และเจอน้ำท่วมขัง มีวิธีการป้องกันและดูแลข้าวให้ปลอดภัยและไม่เสียหายหรือไม่? (Red jasmine rice grown in the South that is flooded; is there a way to protect and care for the rice so that it is safe and not damaged?)

RAG Response: Yes, there are several ways to protect and care for this red sticky rice, such as using good drainage systems and controlling water levels to the right level.

GraphRAG Response: Red jasmine rice planted in the South that is flooded can be protected and cared for safely and without damage by the following methods: 1) Soil preparation before planting, prepare the soil for rice planting by cleaning the soil from weeds and other materials and adding fertilizer to the soil to increase soil moisture and temperature. 2) Planting rice in a position higher than the water level: Plant rice in a position higher than the water level to prevent water from flooding. 3) Using a drainage system: Use an appropriate drainage system to drain water that is stuck on the ground. 4) Caring for rice after planting: Care for rice after planting by cleaning the soil of weeds and other materials and adding fertilizer to the soil to increase soil moisture and temperature. 5) Monitoring for diseases and insects: Monitor for diseases and insects that may occur with rice and use appropriate methods to control diseases and insects. Overall, preventing and caring for red jasmine rice planted in the South that is flooded requires caution and patience in caring for the rice consistently.

Feedback: GraphRAG is better – The user found GraphRAG's response provides more useful tips and methods in a step-by-step format, which makes it more understandable to the user.

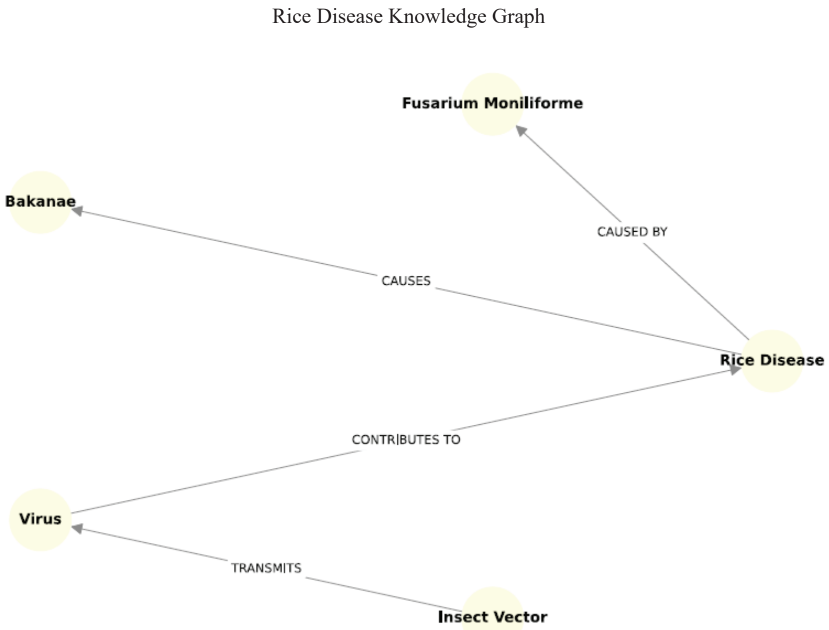

Fig. 5. Rice Disease Knowledge Graph. It visualizes the relationships between different factors related to rice diseases. Entities include types of diseases (e.g., Bakanae, Rice Disease), pathogens (Fusarium Moniliforme, Virus), and vectors (Insect Vector). Relationships depicted are CAUSES, CONTRIBUTES TO, TRANSMITS, and CAUSED BY.

F. Limitation and Future Works

Even though GraphRAG is better than older systems in some ways, it's not perfect. Sometimes, it has trouble with questions that need really deep science knowledge or questions that are unclear and can be understood in different ways. To make it better, we will work on:

Making the Information Base Bigger: We will add more information to the knowledge graph so it knows more things. Getting Better at Understanding Tricky Questions: We will teach the system to be better at figuring out questions that are not clear or have multiple meanings. Improving How Answers Are Made: We will make the system better at creating answers so it can handle many different kinds of questions. Also, we want to use GraphRAG for other types of farming, not just rice. This will help us see if it can work well in different areas [9].

Market Trends and Economics: Questions in this area addressed economic aspects of rice farming, such as:

• What is the current market price for rice?

• What is the demand for organic rice?

Export and Trade Regulations: This category covered the legal and logistical aspects of rice export, including:

• What are the regulations for exporting rice to different countries?

• What is the certification process for exporting organic rice?

IV. System Architecture

The GraphRAG system was developed using a modular architecture to ensure scalability and flexibility in query handling. To effectively compare GraphRAG with a traditional approach, we implemented both GraphRAG and a baseline RAG system, each leveraging distinct data storage and retrieval mechanisms. The system architectures are visualized as shown in Fig. 6 and detailed in the following subsections.

1) *LangChain for Text Extraction* (*Both Systems*): Both the GraphRAG and baseline RAG systems utilize LangChain for extracting relevant textual data from unstructured PDF documents related to Thai rice. Specifically, we employed the *PymuPDF*

module to process PDF documents and the *Langchain core.documents* for chunking text. This ensures consistent and effective text preparation across both systems, focusing the comparative evaluation on the impact of knowledge graph integration [11].

2) *Neo4j Knowledge Graph Database* (*GraphRAG*): The GraphRAG system's core innovation lies in its Neo4j knowledge graph. This graph database is optimized for managing and querying large-scale, interconnected data. Neo4j structures the extracted Thai rice data into entities and relationships, as described previously, enabling GraphRAG to perform graph-based retrieval for contextually rich information. Cypher query language is used for efficient and expressive interaction with the Neo4j database [17].

3) *Pinecone Vector Database* (*Baseline RAG*): In contrast to GraphRAG, the baseline RAG system utilizes Pinecone as its data storage and retrieval mechanism. Pinecone, a vector database, stores vectorized embeddings of the extracted text chunks. For the baseline RAG, we employed sentence-transformers/all-mpnet-base-v2 from Hugging Face embeddings to generate these embeddings. Retrieval in the baseline RAG system is performed using a similarity search over these vector embeddings, a common approach in traditional RAG pipelines.

4) *Llama 3 8b Language Model via Groq API* (*Both Systems*): For response generation in both GraphRAG and the baseline RAG systems, we employed the Meta Llama 3 8b Instruct model, accessed through the Groq API. Utilizing the Groq API provides low-latency inference, crucial for real-time query responsiveness and for facilitating isolating the impact of the knowledge graph versus vector database retrieval on response quality. Specific prompting strategies were designed to guide the LLM to generate informative and domain-specific answers based on theretrieved context [18].

5) *Gradio Interface for preference test and System Evaluation:* To facilitate user evaluation and performance analysis, we developed a Gradio interface. This interface enabled a side-by-side preference test of GraphRAG and the baseline RAG system, allowing participants to directly compare the responses for the same queries. Furthermore, the Gradio interface was instrumented to measure key performance indicators, such as model response time (latency) and memory usage for both systems. These metrics were collected to provide quantitative data on system efficiency alongside qualitative user feedback. The process starts with taking Thai text from PDF files about rice using LangChain. Then, the text is broken into smaller pieces and put into a knowledge graph using Neo4j. When someone asks a question, the system finds the right information from the knowledge graph and sends it to the Llama 3 8b model through the Groq API. Then, the model creates an answer that makes sense with the question.

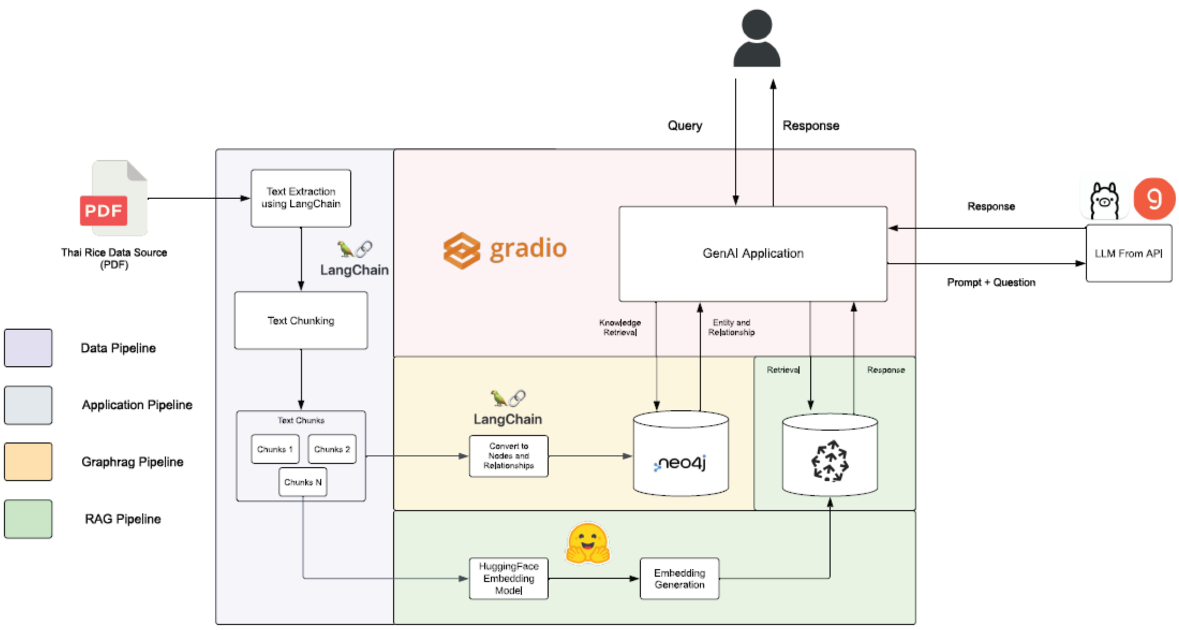

Fig. 6. System Architecture for Thai Rice Assistant. The system is divided into four main pipelines: Data Pipeline, Application Pipeline, GraphRAG Pipeline and RAG Pipeline. Each pipeline is color-coded to highlight its specific role in data processing and retrieval.

V. Results

We performed preference tests on our two different systems using 102 questions. We compared GraphRAG with a basic RAG system. We got feedback from three groups of people: Thai farming experts, researchers studying for their Master's degree at Tuxsa, and AI students from the university. These people helped us see how well GraphRAG worked in real situations.

The answer choices of this test are A was Better, B was Better, both acceptable, or both unacceptable. The results of what they thought are in Table I. This is a blind evaluation. Users will not know the identities of systems A and B. To see how well GraphRAG worked, we tested it against a normal system for getting information called Retrieval-Augmented Generation (RAG). Normal RAG just finds information from messy sources without much understanding. But GraphRAG uses organized links between things like rice types, farming methods, and weather conditions. Because of these connections, they give answers that are more correct and understand the situation better. The results of this experiment are shown in Table I. To evaluate the effectiveness of the proposed GraphRAG system against the baseline RAG, we employed both quantitative performance metrics and qualitative user-based assessment. Performance was measured in terms of average response time and system memory usage during query processing. For qualitative evaluation, we conducted a user preference test where participants rated the relevance, completeness, and overall helpfulness of responses generated by both systems for a predefined set of queries representing common information needs of Thai rice farmers. **While standard metrics like precision and recall are challenging to compute without a large-scale, manually annotated ground-truth dataset for this specific domain, the user preference scores serve as a practical measure of the system's perceived correctness and utility in addressing real-world queries.** Furthermore, the inherent structure of the knowledge graph used in GraphRAG potentially allows for greater explainability compared to standard RAG, as the system can trace information back to specific entities and relationships within the graph [19].

TABLE I
PREFERENCE TEST RESULTS: FEEDBACK ON SYSTEM RESPONSE PREFERENCE

Result	Percentage
RAG	35.3%
GraphRAG	52.9%
Both acceptable	6.9%
Both unacceptable	4.9%

Here are some examples to show why people liked GraphRAG better – because its answers were more useful and right on point:

Question 1: การควบคุมพืชมีวิธีการใดบ้างโดยไม่ต้องใช้สารเคมี? (What are some methods of controlling plants without using chemicals?)

RAG: Good rice seeds will produce good results. If the seeds are not good, have diseases, or are mixed with a lot of weeds, the results will be less.

GraphRAG: Choose good rice varieties with characteristics such as high yield, weather resistance, soil resistance, water resistance, and beauty. The normal RAG answer was not clear. It talked about good seeds but didn't say anything about controlling plants without using chemicals. But the GraphRAG answer was much better. It gave real steps that farmers could take, so it fit the question better. People liked this answer because it was easy to understand and helpful.

Question 2: ข้าวพันธุ์ใดให้ผลผลิตสูงสุด? (Which rice varieties produce the highest yields?)

RAG: Good rice varieties will produce good results because good-quality rice seeds will produce good results.

GraphRAG: Rice variety 105

The normal RAG answer kept saying the same thing over and over and wasn't very specific. But GraphRAG was straight to the point. It said exactly "Rice variety 105", which was a clear and right answer. People liked GraphRAG because it gave direct and useful answers.

User Preference and Feedback

Overall, people thought the GraphRAG answers were better because they were more specific and helpful. Because GraphRAG has a knowledge graph, it can use organized information and connections about Thai rice. This means its answers give you real information that you can use. This shows that GraphRAG is better than normal RAG systems for helping with questions about Thai rice.

TABLE II
AVERAGE RESPONSE TIME AND MEMORY USAGE PER QUESTION

System	Response Time (s)	Memory Usage (KB)
RAG	1.43	457.42
GraphRAG	0.41	213.09

Table II shows the comparison of average response time and memory usage per question between the traditional RAG system and GraphRAG. GraphRAG demonstrates a significant improvement in efficiency, with a response time approximately three times faster than RAG (0.41 seconds compared to 1.43 seconds) and a reduction in memory usage by more than half (213.09 KB compared to 457.42 KB).

The results, particularly the strong user preference for GraphRAG responses and its ability to handle complex queries more effectively than the baseline RAG system, align with the growing body of research highlighting the benefits of incorporating structured knowledge into generative AI systems [20]. Our findings support the hypothesis that leveraging explicit relationship information stored within a knowledge graph allows GraphRAG to provide more contextually relevant and accurate answers, confirming the advantages observed in GraphRAG applications in other domains and demonstrating its effectiveness even within the specific constraints of the low-resource Thai rice information ecosystem.

VI. Conclusion

This research demonstrated the successful application of GraphRAG, a knowledge graph-enhanced RAG approach, to improve query assistance for Thai rice cultivation. The system was developed using a corpus of 6 key documents related to Thai rice farming practices, resulting in a knowledge graph comprising approximately 200 chunks, which consist of 3000 maximum words per chunk. Our comparative analysis showed that GraphRAG significantly outperformed a traditional RAG baseline in user preference tests, particularly for complex queries, while maintaining reasonable performance efficiency. This highlights the value of integrating structured knowledge graphs to enhance contextual understanding and information retrieval accuracy, especially in specialized domains like agriculture.

The tests we did showed that GraphRAG is better than basic systems, especially when dealing with hard questions. These hard questions need a good understanding of how farming works, how rice is sold to other countries, and how different countries compete. By using information from places where there isn't much data, GraphRAG has shown it can be good at finding information even when it's hard to get data.

Future work will focus on several key areas. Firstly, we plan to **expand the knowledge graph** by incorporating a wider range of data sources, including real-time weather data and market prices, to provide more comprehensive and timely information. Secondly, we aim to improve the system's natural language understanding capabilities to better handle ambiguous or colloquial queries from farmers. Thirdly, enhancing the generative component to produce more nuanced and actionable advice is a priority. Finally, exploring the adaptability of this GraphRAG framework to other agricultural domains within Thailand represents a promising avenue for future research.

Acknowledgment

The authors appreciate the support of the Tuxsa Master's Degree Research Group at Thammasat University for helping with data collection and to all participants who contributed critical feedback to this investigation.

References

[1] D. S. Birla, K. Malik, M. Sainger, D. Chaudhary, R. Jaiwal, and P. K. Jaiwal, "Progress and challenges in improving the nutritional quality of rice (Oryza Sativa L.)," *Crit. Rev. Food Sci. Nutr*., vol. 57, no. 11, pp. 2455-2481, Jul. 2017.

[2] R. Wassmann et al., "Climate change affecting rice production: The physiological and agronomic basis for possible adaptation strategies," *Adv. Agron*., vol. 101, pp. 59-122, Oct. 2009.

[3] K. Kiratiratanapruk et al., "Development of paddy rice seed classification process using machine learning techniques for automatic grading machine," *J. Sensors*, vol. 2020, no. 1, p. 7041310, Jul. 2020.

[4] P. P. Bhat, R. Prasad, K. Anil, and A. Jadhav, "The role of information and communication technology in enhancing the effectiveness of agricultural extension programs worldwide: A review," *J. Sci. Res. Rep*., vol. 30, no. 7, pp. 963-976, Jul. 2024.

[5] W. Nwankwo, C. P. Nwankwo, and A. Wilfred, "Leveraging on Artificial Intelligence to accelerate sustainable bioeconomy," *IUP J. Knowl. Manag*. vol. 20, no. 2, pp. 35-59, Apr. 2022.

[6] C. Ling et al., "Domain specialization as the key to make large language models disruptive: A comprehensive survey," *Journal of Economics and Management*, vol. 46, 509-583, Dec. 2023.

[7] Y. Huang and J. Huang, "A survey on retrieval-augmented text generation for large language models." arXiv, 2024. [online]. Available: https://arxiv.org/abs/2404.10981 [Accessed Apr. 2024].

[8] S. Rezayi et al., "Exploring new frontiers in agricultural NLP: Investigating the potential of large language models for food applications," IEEE Trans. Big Data, 2024.

[9] Q. Zhang et al., "A survey of graph retrieval-augmented generation for customized large language models." arXiv, 2025. [online]. Available: https://arxiv.org/abs/2501.13958 [Accessed Jan. 2025].

[10] B. Peng et al., "Graph retrieval-augmented generation: A survey." arXiv, 2024. [online]. Available: https://arxiv.org/abs/2408.08921 [Accessed Aug. 2024].

[11] S. Ghane, R. Sawant, G. Supe, and C. Pichad, "LangchainIQ: Intelligent content and query processing," *Int. J. Manag. Technol. Soc. Sci. (IJMTS)*, vol. 9, no. 3, pp. 34-43, 2024.

[12] ArangoDB, "ArangoDB's GraphRAG Transforms Healthcare Data Management," *ArangoDB Blog*, 2025. [online]. Available: https://arangodb.com/2025/04/arangodbs-graphrag-transforms-healthcare-data-management/ [Accessed Apr. 6, 2025].

[13] T. Balarabe. "GraphRAG vs Traditional RAG: Knowledge Graphs for Accurate, Enhanced RAG Applications," Medium.com, 2025. [online]. Available: https://medium.com/@tahirbalarabe2/graphrag-vs-traditional-rag-knowledge-graphs-for-accurate-enhanced-rag-applications-2cc4f6f9f4b4 [Accessed: Apr. 6, 2025].

[14] A. N. Arifin, "Are Re-Ranking in retrieval-augmented generation methods impactful for small agriculture qa datasets? a small experiment," *BIO Web of Conf*., vol. 167, pp. 1-6, Mar. 2025, https://doi.org/10.1051/bioconf/202516701001

[15] C. Pechsiri and R. Piriyakul, "Developing the UCKG-WHY-QA system," in *Proc. 7th Int. Conf. Comput. Converg. Technol. (ICCCT)*, *IEEE*, 2012, pp. 679-683.

[16] Y. Xie, L. Jia, and J. Dai, "Construction of a traditional Chinese medicine Dao Yin science knowledge graph based on Neo4j," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, *IEEE*, 2023, pp. 4662-4666.

[17] C. Johnpaul and T. Mathew, "A Cypher query based NoSQL data mining on protein datasets using Neo4j graph database," in *Proc. 4th Int. Conf. Adv. Comput. Commun. Syst. (ICACCS), IEEE*, 2017, pp. 1-6.

[18] R. Sapkota et al., "Multi-modal LLMs in agriculture: A comprehensive review," *TechRxiv*. [Online]. Available: https://www.techrxiv.org/users/795986/articles/1224079-multi-modal-llms-in-agriculture-a-comprehensive-review [Accessed Sep. 2024].

[19] Lettria, "10 rules for optimizing your GraphRAG strategies," *Lettria Blog*, 2025. [Online]. Available: https://www.lettria.com/blogpost/10-rules-for-optimizing-your-graphrag-strategies [Accessed Apr. 6, 2025].

[20] AWS, "Improving Retrieval Augmented Generation accuracy with GraphRAG," *AWS Machine Learning Blog*, 2025. [Online]. Available: https://aws.amazon.com/blogs/machine-learning/improving-retrieval-augmented-generation-accuracy-with-graphrag/ [Accessed Apr. 6, 2025].

Gampanut Soontontam is currently a student in the Bachelor of Computer Engineering and Artificial Intelligence at Panyapiwat Institute of Management, His research interests are in the area of Machine Learning and Large Language Models.

Tinnaphob Dindam is an Assistant Professor in Electrical Engineering and Lecturer in Computer Engineering and Artificial Intelligence at Panyapiwat Institute of Management, Nonthaburi, Thailand.
He graduated with a B.S.Tech. Ed., M.Eng., and Ph.D. in Electrical Engineering from King Mongkut's University of Technology North Bangkok. His research focuses on wireless networks, cognitive radio, and Machine Learning.

Adisorn Kheaksong is an Assistant Professor in Computer Engineering and Lecturer in Computer Engineering and Artificial Intelligence at Panyapiwat Institute of Management, Nonthaburi, Thailand.
He received his B.Ind. Eng. in Electrical Engineering from Saint John University, M.Sc. in Computer Science from Rangsit University, and Ph.D. in Electrical Engineering from King Mongkut's University of Technology North Bangkok. His research focuses on wireless networks, cloud computing, and emerging technologies in artificial intelligence.

Kanabadee Srisomboon is an Associate Professor of the Department of Electrical and Computer Engineering at King Mongkut's University of Technology North Bangkok, Bangkok, Thailand. He received his B.S. degree from Civil Aviation Training Center in 2011. The M.Eng. and Ph.D. degrees from King Mongkut's University of Technology North Bangkok were received in 2013 and 2016, respectively. His research interests include image processing, smart grid communication, Cognitive Radio Communication, and Artificial Intelligence technology.

Parinya Sanguansat is an Associate Professor in Electrical Engineering and head of Computer Engineering and Artificial Intelligence at Panyapiwat Institute of Management, Nonthaburi, Thailand.
He graduated B.Eng,. M.Eng. and Ph.D. in Electrical Engineering from Chulalongkorn University. His research areas include Machine Learning, Computer Vision, and Artificial Intelligence.

Effect of Packaging Materials on the Quality and Shelf Life of Fresh-Cut Vegetables

Chairat Burana[1*], Phatcharee Kittisuban[2], Ruamporn Liamkaew[3], and ENDO Gen[4]
[1]Faculty of Innovative Agriculture and Management, Panyapiwat Institute of Management, Nonthaburi, Thailand
[2]Faculty of Food Science, Technology and Management, Panyapiwat Institute of Management, Nonthaburi, Thailand
[3]Faculty of Food Business Management, Panyapiwat Institute of Management, Nonthaburi, Thailand
[4]Faculty of International Relations, Daito Bunka University, Saitama, Japan
E-mail: chairatbur@pim.ac.th*, phatchareekit@cpram.co.th, ruampornlia@pim.ac.th, endogen@ic.daito.ac.jp

Received: April 9, 2025 / Revised: May 15, 2025 / Accepted: May 19, 2025

Abstract —**The fresh-cut products market has seen significant growth over the past decade, expanding from foodservice to retail shelves, convenience stores, and mobile fruit vans. Controlling temperature, atmosphere, relative humidity, and sanitation is crucial for maintaining the quality, safety, and shelf life of fresh-cut produce. The effect of different types of packaging on the quality and shelf life of fresh-cut lettuce, onion, and mixed cabbage with carrot was investigated. All fresh-cut produce was packaged in three types of packaging (A-Best®, Fresh & Fresh®, Active Pak®, and A-Best® at 5°C) and then stored at 4°C to simulate the refrigerated shelf in convenience stores. The result found that fresh-cut lettuce, onion, and mixed cabbage with carrot in A-Best at 5°C showed the highest overall acceptability both externally and internally, significantly. On the other hand, A-Best® and A-Best® at 5°C effectively retarded the rate of browning symptom of the cut surfaces. The results suggest that combined treatments showed better results than those in the single treatment and have commercial potential in improving the shelf life and maintaining the quality of fresh-cut produce.**

Index Terms — **Quality Control, Fresh-Cut, Postharvest Processing, Packaging, Modified Atmosphere**

I. Introduction

Fresh-cut produce is commonly sold in open-air markets and food stalls across many Asian countries and is increasingly available in supermarkets. In particular, fresh-cut fruits have grown in popularity in the region's urban centers. When displayed without refrigeration, these products generally do not remain viable beyond the day of display.

Fresh-cut vegetables command a greater market share than fresh-cut fruits in Thailand [1]. As consumer demand for ready-to-eat products increases, the fresh-cut market in Thailand is projected to experience sustained growth. The range of fresh-cut vegetables has expanded considerably, including items such as carrots, celery, broccoli, lettuce, and various salad mixes. Innovations in packaging and processing have improved shelf life and quality, making these products more appealing to consumers. For example, the development of modified atmosphere packaging helps maintain freshness by controlling oxygen (O_2) and carbon dioxide (CO_2) gas within the package [2].

Fresh-cut produce generally exhibits higher respiration rates than whole, unprocessed produce. This increased respiration accelerates metabolic activity, leading to rapid senescence. In addition, a higher respiration rate contributes to an accelerated loss of flavor, comprising acids, sugars, and other important elements, as well as a decline in nutritional value. Moreover, ensuring the quality and nutritional stability of fresh-cut produce during storage is more demanding than for its whole counterparts [3].

Packaging is effective in maintaining the quality and extending the shelf life of fresh-cut vegetables by reducing O_2 and increasing CO_2 in the package [4]. Modified Atmosphere Packaging (MAP) for fresh products involves adjusting the internal atmosphere of the package to optimize product preservation. This is achieved through the natural interaction between the product's respiration rate and the gas exchange properties of the packaging material. MAP can be implemented using either active or passive methods [5], [6]. However, postharvest problems of packaged fresh-cut lettuce, onion, and mixed cabbage with carrot in the retail or convenience store include browning, physiological disorders, non-optimal storage conditions, and inappropriate packaging materials.

Fig. 1. Display of fresh-cut products in a modern trade in Thailand [7].

II. Research Objective

The present study aims to investigate the effect of packaging materials on the quality and shelf life of fresh-cut products.

III. Materials and Methods

A. Plant Materials and Treatment Conditions

Plant Materials

Fresh lettuce, onion, cabbage, and carrot were purchased from the wholesale market in Pathum Thani province, Thailand. After purchase, all commodities were brought to the laboratory at the School of Bioresources and Technology, King Mongkut's University of Technology Thonburi (Bang Khun Thian) within 2 hours of transportation.

Fresh-Cut Products are Prepared for Processing

Fresh-cut products were prepared by longitudinal slicing using a sharp knife. Then, all fresh-cut vegetables were immersed in 2% calcium chloride ($CaCl_2$) solution as a preservative and curing agent at 25°C for 15 minutes. Then, all fresh-cut were controlled in a drain basket, dried, and 100 grams each were put into different packaging as follows: A-Best®, Fresh & Fresh®, Active Pak® at 25°C, and A-Best® with 5°C (Applying cold water at 5°C). After packaging, all treatments were stored at a constant temperature of 4°C and 65-80% relative humidity (RH) until the end of the experimental period.

B. Gas Composition Analyses

At each sampling point, the levels of O_2 and CO_2 accumulation within the packages of fresh-cut were monitored. Gas measurements were conducted using a PBI-Dansensor CheckMate II (Denmark) by inserting the instrument's needle directly into the packaging.

C. Sensory Acceptance

Overall acceptability was rated by an untrained panel of 15 judges using a 9-point hedonic scale [8], with number 9 = "like very much" and number 1 = "dislike very much".

D. Statistical Analysis

All statistical analyses were conducted using JMP statistical software (SAS Institute Inc., Cary, NC, USA). Data were analyzed using Student's t-test and Analysis of Variance (ANOVA). When ANOVA indicated significant differences ($P < 0.05$), mean separation was performed using the Tukey-Kramer test.

IV. Results and Discussion

A. Concentration of O_2 in Fresh-Cut Packaged

The O_2 concentration in the A-Best® and A-Best® with 5°C package of lettuce (Fig. 2), onion (Fig. 3), and mixed cabbage with carrot (data not shown) gradually decreased throughout the experimental period. While in Fresh & Fresh®, Active Pak® resulted in slightly decreased O_2 concentration and higher than that in A-Best® and A-Best® with 5°C package (Fig. 1 and Fig. 2). Low oxygen levels are a key strategy for managing the shelf life of fresh produce in MAP. In early MAP applications, the main objective of reducing O_2 was to slow the respiration rate of fruits and vegetables, thereby preserving quality and extending shelf life [9]. However, low O_2 also suppresses the growth of aerobic microorganisms [10].

Recent studies have indicated that low O_2 conditions may inhibit the development of desirable flavors in fresh melons [11] or contribute to undesirable flavors in baby spinach [12]. This suggests that new technologies utilizing low O_2 levels should be tailored to the specific characteristics of each type of produce to avoid triggering anaerobic respiration.

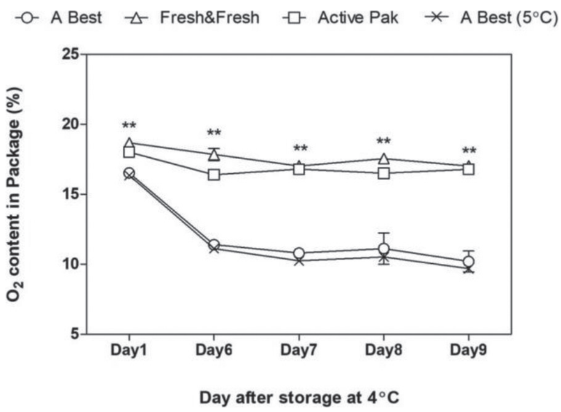

Fig. 2. Levels of O_2 concentration in fresh-cut lettuce packed in different material packages and stored at 4°C for 9 days. Each symbol is the mean of three replicate measurements; vertical lines represent SE. (○) A-Best®, (Δ) Fresh & Fresh®, (□) Active Pak®, and (*) A-Best® at 5°C

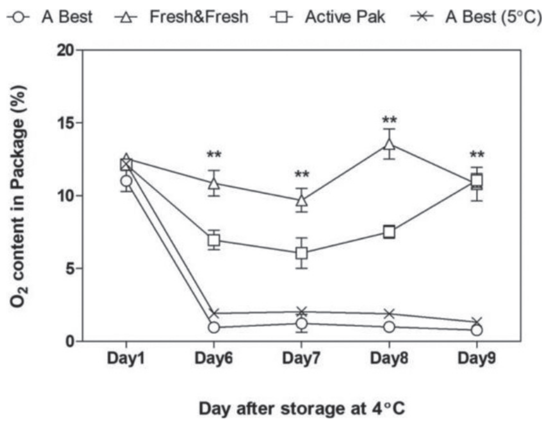

Fig. 3. Levels of O_2 concentration in fresh-cut mixed cabbage with carrot packed in different material packages and stored at 4°C for 9 days. Each symbol is the mean of three replicate measurements; vertical lines represent SE. (○) A-Best®, (Δ) Fresh & Fresh®, (□) Active Pak®, and (*) A-Best® at 5°C

- *Concentration of CO_2 in Fresh-Cut Packaged*

The carbon dioxide concentration in the package of fresh-cut lettuce was about 4%, then it decreased slightly until the end of the observation, and there was no significant difference between the package A Best® and A-Best® at 5°C. However, the lower CO_2 concentration of 1.5-2% was found in Fresh & Fresh® and Active Pak® (Fig. 4). Similarly with lettuce, the CO_2 concentration in fresh-cut mixed cabbage with carrot showed higher in A Best® and A-Best® with 5°C but the difference between treatments was significant (data not shown).

Overall, there are many similarities between the effects of low O_2 and high CO_2 on the suppression of various metabolic processes in fruits and vegetables [13]. Respiration is generally inhibited by low O_2 and high CO_2 environments [14].

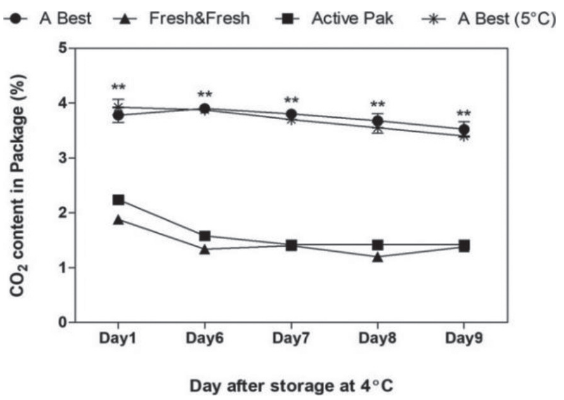

Fig. 4. Levels of CO_2 concentration in fresh-cut lettuce packaged in different material packaging and stored at 4°C for 9 days. Each symbol is the mean of three replicate measurements; vertical lines represent SE. (○) A Best®, (Δ) Fresh & Fresh®, (□) Active Pak®, and (*) A-Best® at 5°C

Initially, the respiration rate of the product is much higher than the package permeation rates for CO_2 and O_2. This results in the accumulation of CO_2 and depletion of O_2 within the package [10].

Reduced O_2 and increased CO_2 concentrations must be stringent enough to slow metabolism, resulting in reduced chlorophyll loss along with decreased pigment accumulation and prolonged shelf-life [15].

- *Overall Acceptance of Fresh-Cut*

According to the consumer survey, consistent appearance, freshness, and aroma are important quality attributes for fresh-cut lettuce, onions, mixed cabbage, and carrots. According to several researchers [16], [17], in addition to a fresh appearance, an acceptable texture of the fresh-cut items and nutritional value are also important.

The initial scores for overall acceptability outside of fresh-cut lettuce (Fig. 5), onion (data not shown), and mixed cabbage with carrot (data not shown) on a scale of 1-9 were 9 for all treatment materials. The scores for A-Best® and Active Pak® drastically decreased to 6.3 and 6.2 after 6 days of storage. For lettuce, fresh-cut lettuce wrapped in A-Best® at 5°C had the highest final score on the last day of storage with 6, followed by A-Best®, Fresh & Fresh®, and Active Pak® (Fig. 5).

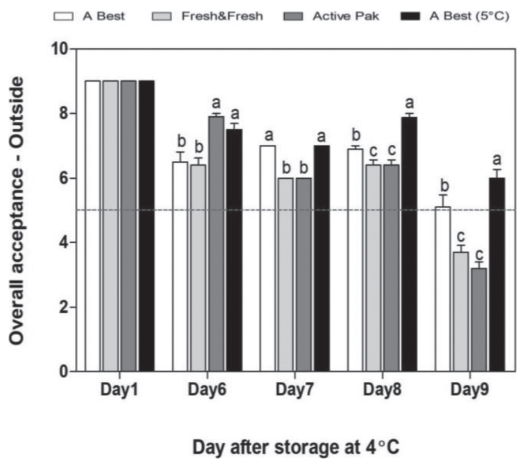

Fig. 5. Total acceptance both outside in fresh-cut lettuce packed in different material packaging and stored at 4°C for 9 days.

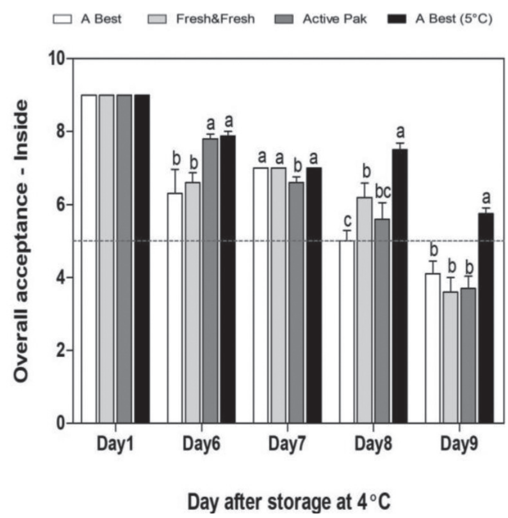

Fig. 6. Total acceptance both inside ~~in~~ fresh-cut lettuce packed in different material packaging and stored at 4°C for 9 days.

Similarly, with the overall acceptability outside, overall acceptability inside, A-Best® with 5°C shows earned the highest final scores on the last day of storage with 5.8, but no difference between treatment and all other treatments (Fig. 6).

Fig. 7. Overall appearance of fresh-cut onions packaged in A Best®, Fresh & Fresh®, Active Pak®, and A-Best® at 5°C and stored at 4°C for 9 days.

Fig. 8. Overall appearance of fresh-cut onions packaged in A Best®, Fresh & Fresh®, Active Pak®, and A-Best® at 5°C and stored at 4°C for 9 days.

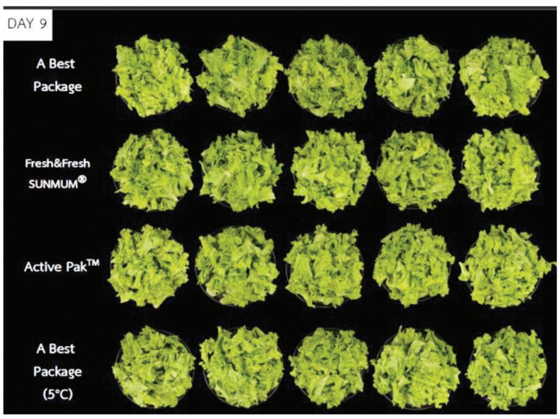

Fig. 9. Overall appearance of fresh-cut lettuce packaged in A Best®, Fresh & Fresh®, Active Pak®, and A-Best® at 5°C and stored at 4°C for 9 days.

V. Conclusion

The quality of fresh-cut lettuce, onion, and mixed cabbage with carrot is determined by a combination of attributes, properties, and characteristics that influence consumer value. Among these, "appearance" and "freshness" have been identified as the most important factors at the point of purchase. This study evaluated the effectiveness of three packaging materials in preserving these attributes in fresh-cut produce. The results showed that A-Best® and A-Best® at 5°C were the most effective in maintaining overall quality, including appearance, cut surface color, and freshness. Future research will explore the impact of adding antioxidants, fungicides, and vitamins to improve the nutritional value of fresh-cut produce.

References

[1] K. Sa-nguanpuag, S. Kanlayanarat, and K. Tanprasert, "Trends of fresh-cut produce in Thai retail markets for identification of packaging for shredded green papaya," *Acta Horticulturae*, vol. 746, pp. 481-483, Aug. 2007.

[2] U. De Corato, "Improving the shelf-life and quality of fresh and minimally-processed fruits and vegetables for a modern food industry: A comprehensive critical review from the traditional technologies into the most promising advancements," *Crit. Rev. Food Sci. Nutr.*, vol. 60, no. 6, pp. 940-975, Jan. 2019.

[3] M. Cantwell and T. Suslow, "Fresh-cut fruits and vegetables: Aspects of physiology, preparation and handling that affect quality," in *Annual Workshop Fresh-Cut Products: Maintaining Quality and Safety*, vol. 5, CA, USA: Univ. of California, 1999, pp. 1-2.

[4] Y. Makino, S. Oshita, Y. Kawagoe, and A. Tanaka, "Simultaneous prediction of Oxygen and Carbon Dioxide concentrations in a perforated pouch with light red tomato fruits by a mathematical model," *Trans. ASABE*, vol. 51, no. 2, pp. 559-565, Feb. 2007.

[5] L. Lee, J. Arul, R. Lencki, and F. Castaigne, "A review on modified atmosphere packaging and preservation of fresh fruits and vegetables: Physiological basis and practical aspects-Part I," *Packag. Technol. Sci.*, vol. 8, no. 6, pp. 315-331, Nov. 1995.

[6] R. M. Beaudry, "Effect of O_2 and CO_2 partial pressure on selected phenomena affecting fruit and vegetable quality," *Postharvest Biol. Technol.*, vol. 15, no. 3, pp. 293-303, Mar. 1999, https://doi: 10.1016/S0925-5214(98)00092-1

[7] C. Burana, N. Sirisukchaitavorn, and I. M. Rumainum, "Post-harvest treatment to prolong the shelf life and maintain the quality of fresh-cut produce," in *Proc. 14th Nat. and 10th Int. PIM Conf.*, 2024, pp. 482-489.

[8] E. Larmond, *Laboratory Methods for Sensory Evaluation of Food*, Ottawa, ON, Canada: Canadian Government Publishing Centre, 1977, p. 44.

[9] M. D. Wilson, R. A. Stanley, A. Eyles, and T. Ross, "Innovative processes and technologies for modified atmosphere packaging of fresh and fresh-cut fruits and vegetables," *Crit. Rev. Food Sci. Nutr.*, vol. 59, no. 3, pp. 411-422, Sep. 2017, https://doi.org/10.1080/10408398.2017.1375892.

[10] A. A. Kader, D. Zagory, E. L. Kerbel, and C. Y. Wang, "Modified atmosphere packaging of fruits and vegetables," *Crit. Rev. Food Sci. Nutr.*, vol. 28, pp. 1-30, Sep. 2009, https://doi.org/10.1080/10408398909527490

[11] A. L. Amaro, J. C. Beaulieu, C. C. Grimm, R. E. Stein, and D. P. F. Almeida, "Effect of oxygen on aroma volatiles and quality of fresh-cut cantaloupe and honeydew melons," *Food Chem.*, vol. 130, no. 1, pp. 49-57, Jan. 2012, https://doi.org/10.1016/j.foodchem.2011.06.052

[12] J. A. Tudela, A. Marín, Y. Garrido, M. Cantwell, M. S. Medina-Martínez, and M. I. Gil, "Off-odour development in modified atmosphere packaged baby spinach is an unresolved problem," *Postharvest Biol. Technol.*, vol. 75, pp. 75-85, Jan. 2013, https://doi.org/10.1016/j.postharvbio.2012.08.006

[13] F. M. Mathooko, "Regulation of respiratory metabolism in fruits and vegetables by carbon dioxide," *Postharvest Biol. Technol.*, vol. 9, no. 3, pp. 247-264, Dec. 1996, https://doi.org/10.1016/S0925-5214(96)00019-1

[14] J. Mattheis and J. K. Fellman, "Impacts of modified atmosphere packaging and controlled atmospheres on aroma, flavor, and quality of horticultural commodities," *HortTechnology*, vol. 10, no. 3, pp. 507-510, Jul. 2000, https://doi.org/10.21273/HORTTECH.10.3.507

[15] A. Watada, N. Ko, and D. Minott, "Factors affecting quality of fresh-cut horticultural products," *Postharvest Biol. Technol.*, vol. 9, no. 2, pp. 115-125, Nov. 1996, https://doi.org/10.1016/S0925-5214(96)00041-5

[16] V. D. Schlimme, "Marketing lightly processed fruits and vegetables," *Hortic. Sci.*, vol. 30, no. 1, pp. 15-17, Feb. 1995, https://doi.org/10.21273/HORTSCI.30.1.15

Chairat Burana was born on 5th June 1983 in Sisaket Province. He received a Bachelor of Science in Agricultural Science from Mahidol University, Thailand. He obtained his Master of Science in Postharvest Technology international program from King Mongkut's University of Technology, Thonburi, Bangkok, Thailand. During his Master's, he was a research exchange student at a laboratory of horticulture, the Faculty of Agriculture, Utsunomiya University, which was supported by Japan Student Services Organization (JASSO) Scholarship. In 2013, he presented his work at the national and international conferences. Additionally, He has received a research award called "Dr. Adel A. Kader Award for Young Scientists" at XI Controlled & Modified Atmosphere Research International Conference in 2013, held in Trani, Italy. He received his Doctor of Philosophy (Ph.D.) in Bio-production Science from Tokyo University of Agriculture and Technology, Tokyo, Japan, in 2014. During his Ph.D. course, he was supported by the Japanese Government (MEXT) Scholarship and the Tsuji Asia Foundation Scholarship. His research interests are Postharvest Technology: Maintaining the qualities, deterioration, and handling systems for Agricultural products. In the present (2025), he is an Assistant Dean for Research in the Faculty of Innovative Agriculture and Management, Panyapiwat Institute of Management.

Phatcharee Kittisuban is a lecturer of the Faculty of Food Science, Technology and Management, Panyapiwat Institute of Management, and the Head of Research Center (Assistant General Manager) CP Foodlab Co., Ltd., located at Food Innopolis Project Thailand Science Park (NSTDA), which serves the knowledge and research communities for R&D of ready-to-eat food products in CP Group, as well as national and international food business companies. She received the Bachelor of Science (1st class honors; Food science and technology), Mahidol University, Thailand, 2003-2007. She obtained a Doctor of Philosophy (Biotechnology), Major Food Biotechnology, Mahidol University, 2007-2014 (Scholarship: The Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program). She is a founder of CP Foodlab Co., Ltd., located at Food Innopolis Project Thailand Science Park (NSTDA), 2016.

Ruamporn Liamkaew is a lecturer at the Faculty of Food Business Management, Panyapiwat Institute of Management. She received the Bachelor of Science in Biotechnology from Thammasat University, Thailand, in 2005. She obtained her Master of Engineering (2007) and Doctor of Engineering (2014) in Food Engineering from King Mongkut's University of Technology Thonburi, Thailand. She received a scholarship from the Thailand Research Fund through the Royal Golden Jubilee Ph.D. Program: RGJ- industry program, and she was a visiting researcher at the Department of Biophysics, Syracuse University, New York, USA, from 2009 to 2010. She received a research award (bronze award) from Thailand Research Expo 2020 for the research project on potential development for the production and marketing of buffalo milk in Chachoengsao Province. In 2021, she also received a new researcher award at a very good level from the Association of Private Higher Education of Thailand under the patronage of her Royal Highness, Mahachakri Sirindhorn. Her research interests are food product and process development, shelf life extension and evaluation, sensory evaluation, and consumer testing. Currently, she is a deputy dean for academic affairs, faculty of food business management, Panyapiwat Institute of Management.

ENDO Gen is a professor at the Faculty of International Relations and the Graduate School of Asian Area Studies at Daito Bunka University, Japan. He was born on April 18, 1963, in Osaka, Japan. He received his Ph.D. in Science from the University of Tokyo, Japan, in 1998. While a Ph.D. student at the University of Tokyo, he received a scholarship from the Osaka International House Foundation and the Matsushita International Foundation to study at Chiang Mai University, Thailand, as a graduate student and visiting researcher from 1992 to 1995, where he studied family business and the distribution industry in northern Thailand. He received the academic award of the Business History Society of Japan in 2002 for his work on the family business in northern Thailand. His current research interests include the impact of the expansion of modern retail chains on Thailand's agricultural production and distribution system and the development of fair relationships between small producers and consumers. He is currently Head of the Department of International Relations in the Faculty of International Relations at Daito Bunka University.

Intelligent Assessment of Athlete Physical Fitness: Addressing Data Imbalance

Janyarat Phrueksanant[1*], Chayanont Awikunprasert[2], Jirachai Karawa[3], and Sutthirak Wisetsang[4]
[1]Department of Information Technology and Computer Innovation,
Faculty of Management Sciences and Information Technology, Nakhon Phanom University, Nakhon Phanom, Thailand
[2,3,4]Department of Sports Science, Faculty of Management Sciences and Information Technology,
Nakhon Phanom University, Nakhon Phanom, Thailand
E-mail: janyarat@npu.ac.th*, chayanona@yahoo.com, jckwmosza@gmail.com, watini8939@gmail.com

Received: February 28, 2025 / Revised: May 28, 2025 / Accepted: May 29, 2025

***Abstract*— This study aims to mitigate the impact of imbalanced data through the use of the oversampling technique and to develop supervised learning models for assessing the physical fitness of youth athletes. The dataset comprises the physical fitness test results from 75 athletes aged 11 to 16 years. The dataset presents two major challenges: A limited sample size and a significant class imbalance, with certain fitness levels being underrepresented. This class imbalance can substantially degrade the performance of classification models, as it often leads to biased predictions favoring the majority class while failing to learn the characteristics of minority classes, those that may be most critical in practice.**

To address this issue, the Synthetic Minority Oversampling Technique (SMOTE) was employed to synthetically balance the class distribution. Five supervised learning algorithms were evaluated: Light Gradient Boosting Machine, Decision Tree, Random Forest, Neural Network, and Multinomial Logistic Regression. The Light Gradient Boosting Machine model yielded the highest accuracy at 87.76%, followed by Decision Tree, Random Forest, and Neural Network models, each with an accuracy of 79.59%. The Multinomial Logistic Regression model achieved the lowest accuracy at 75.51%. On average, the classification accuracy across all models improved to 81.41%, representing a 12.23% increase compared to using the original imbalanced dataset. The results demonstrate that applying oversampling techniques such as SMOTE can effectively alleviate the effects of class imbalance and enhance the predictive performance of machine learning models in the context of physical fitness assessment.

***Index Terms*—Athletes, Data Imbalance, Physical Fitness Performance, SMOTE, Supervised Learning**

I. Introduction

In recent years, the integration of data analytics into sports has advanced significantly, enabling more informed decision-making for both individual athletes and teams. Two case studies highlight this trend. The first involves a trainer for a women's college soccer team who utilizes wearable devices to collect internal (e.g., heart rate, body temperature, respiration) and external load data (e.g., running distance, speed, acceleration). These insights are used to monitor training intensity, prevent injuries, and ensure players are physically prepared for competition. Additional tools, such as single-leg squat tests, concussion assessments, sleep tracking, and periodic brain scans, contribute to a comprehensive understanding of each athlete's condition. The second case centers on a college football team led by a head coach and supported by a team operations expert. This team leverages annotated game footage, decision tree models, heatmaps, and time-series analytics to study opponents' tactics and optimize in-game strategy. Data-driven analysis is also applied to recruitment, incorporating advanced performance metrics such as reaction time, spatial awareness, and route-running precision, moving beyond traditional physical measurements. These examples demonstrate how data-driven approaches are transforming sports by enhancing athletic performance, minimizing risk, and informing strategic planning at both individual and organizational levels [1].

Physical fitness is a vital component of health and well-being, serving as the foundation for daily activities [2], [3]. Enhancing physical fitness not only improves quality of life but also serves as a key indicator of physical development over time. In athletic contexts, fitness assessments are commonly used to evaluate health status, identify individual strengths and weaknesses, and guide personalized training strategies aimed at optimizing performance.

This study aims to develop supervised learning models for assessing physical fitness levels among athletes. The dataset used was collected from physical fitness tests conducted at Nakhon Phanom Sports School, Thailand. As is common in real-world datasets, the collected data in this study are imbalanced, with one or more classes significantly underrepresented. Class imbalance is a well-documented challenge in data analytics, particularly in classification tasks, as it can negatively impact the accuracy and generalizability of predictive models. A relevant example is a previous study that utilized real-world data consisting of a small and imbalanced dataset to explore academic performance among IT students. The study employed Principal Component Analysis (PCA) and clustering techniques, and despite having only 115 samples, it successfully extracted meaningful insights. This highlights the effectiveness of dimensionality reduction and unsupervised learning methods in constrained data environments [4]. Given that most supervised learning algorithms assume balanced class distributions, several techniques have been proposed to address this issue [5]-[7]. In this study, the Synthetic Minority Oversampling Technique (SMOTE) is applied to augment the minority class, aiming to improve model robustness and classification performance.

The remainder of this paper is structured as follows. Section II provides a review of supervised learning techniques, the Synthetic Minority Oversampling Technique (SMOTE), and related literature. Section III describes the dataset, outlines the preprocessing procedures, and explains the implementation of SMOTE, as well as the models and evaluation metrics employed in the study. Section IV presents and analyzes the experimental results. Finally, Section V concludes the paper by summarizing the key findings and suggesting directions for future research.

II. Literature Review

A. *State-of-the-Art Supervised Learning Techniques for Classification*

Classification is one of the most extensively explored tasks in data analytics, aiming to categorize input data into predefined classes or labels. Supervised learning techniques form the foundation of classification models by learning patterns from labeled training data. Traditional methods such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) have been widely adopted due to their interpretability and effectiveness on structured datasets. More recent advancements include Ensemble Methods like Random Forests and Gradient Boosting Machines (e.g., XGBoost, LightGBM), which combine multiple weak learners to achieve higher accuracy and robustness. In addition, Artificial Neural Networks (ANNs) and deep learning architectures, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), have demonstrated exceptional performance in complex, high-dimensional data such as images and sequences. These state-of-the-art models offer improved predictive power but often require larger datasets and computational resources. The selection of an appropriate classifier depends on factors such as data characteristics, model interpretability, scalability, and computational constraints. This section provides an overview of commonly used techniques, highlighting their strengths and state-of-the-art approaches.

1) Traditional Machine Learning Techniques

Logistic Regression (LR): Logistic Regression is a simple yet effective method for binary classification. It models the probability of a data point belonging to a class using a logistic function. Despite its simplicity, it performs well with linearly separable data and is widely used in various fields such as healthcare and finance.

Support Vector Machines (SVM): SVM is a robust classifier that aims to find the optimal hyperplane that separates data points of different classes with the maximum margin. Kernel methods allow SVM to handle non-linear data by transforming it into higher-dimensional spaces.

Decision Trees (DT): Decision Trees are hierarchical models that split data into subsets based on feature thresholds. They are interpretable and can handle both categorical and numerical data.

Ensemble Methods: Ensemble methods like Random Forest and Gradient Boosting (e.g., XGBoost, LightGBM, CatBoost) combine multiple weak learners to improve predictive performance.

2) Neural Networks and Deep Learning

Artificial Neural Networks (ANNs): ANNs are the foundation of modern deep learning. They consist of layers of interconnected nodes (neurons) that learn hierarchical representations of input data. ANNs are versatile and can model complex, non-linear relationships.

Convolutional Neural Networks (CNNs): CNNs are specialized for image and spatial data classification. They use convolutional layers to automatically extract features from raw data, making them state-of-the-art for tasks like object detection and image recognition.

Recurrent Neural Networks (RNNs) and Variants: RNNs, including Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU), are used for sequential data like text, speech, and time series. They capture temporal dependencies in data.

3) Emerging Techniques and Trends

Explainable AI (XAI) in Classification: Modern classification tasks increasingly focus on interpretability. Techniques like SHAP (Shapley Additive Explanations) and LIME (Local Interpretable Model-agnostic Explanations) help demystify complex models like deep learning.

Hybrid Models: Combining traditional machine learning with deep learning is becoming a trend, such as integrating feature engineering from domain knowledge into neural network architectures.

Supervised learning techniques for classification have evolved significantly, from traditional algorithms like logistic regression and support vector machines to advanced methods involving neural networks and transformers. Each technique has its strengths and limitations, making them suitable for specific types of data and tasks. Emerging trends, such as explainable AI, few-shot learning, and hybrid models, are paving the way for more robust and interpretable classification systems. As data continues to grow in complexity and volume, leveraging the right combination of techniques will be critical to advancing the field and solving real-world problems [8]-[12].

B. *Synthetic Minority Oversampling Technique*

Synthetic Minority Oversampling Technique (SMOTE) is a popular method for addressing class imbalance in datasets. It was introduced by Chawla et al. in 2002 [13]. The technique works by creating synthetic examples in the feature space to increase the representation of the minority class, rather than duplicating existing instances. For each sample in the minority class, it identifies its k-nearest neighbors within the minority class. Typically, k=5 is used. Then a random neighbor is chosen from the k-nearest neighbors. A new synthetic sample is created by interpolating between the original sample and the chosen neighbor. The formula is:

$$\mathcal{X}_{synthetic} = \mathcal{X}_i + (\mathcal{X}_j - \mathcal{X}_i).\,\lambda \quad (1)$$

Where:

$\mathcal{X}_i$ is the original minority class sample.

$\mathcal{X}_j$ is the chosen neighbor.

λ is a random number between 0 and 1.

This approach helps improve the performance of machine learning models, particularly those sensitive to class imbalance, like decision trees and neural networks.

C. *Related Work*

Studies on imbalanced data classification span diverse domains, employing advanced techniques to address challenges associated with class imbalance. One study focused on predicting osteoarthritis conditions in elderly patients, utilizing a dataset of 370 records divided into four classes: No symptoms, early symptoms, moderate symptoms, and severe symptoms. The study applied ADASYN and SMOTE to balance the data, using a 10-fold cross-validation method along with one-vs-one and one-vs-all multi-class classification approaches. The combination of ADASYN and the one-vs-one method achieved an impressive accuracy of 97.31% [14]. Another study examined career path predictions for 2,005 university graduates using employment status data. By leveraging oversampling, undersampling, and SMOTE, alongside decision tree and random forest classifiers, the study found that random forests, applied to oversampled data, yielded the best performance, achieving 67.17% accuracy, with corresponding precision, recall, and F-measure values of 0.66, 0.67, and 0.66, respectively [15]. Another interesting study that addresses the challenge of imbalanced datasets in predicting first-year engineering students' performance employs oversampling methods—SMOTE, Borderline-SMOTE, SVMSMOTE, and ADASYN—to balance the data, followed by classification using models such as Multi-Layer Perceptron (MLP), Gradient Boosting, AdaBoost, and Random Forest. The findings indicate that combining Borderline-SMOTE with various classifiers enhances the prediction accuracy for minority classes, thereby aiding in the early identification of students at risk of underperformance. This research underscores the effectiveness of integrating oversampling techniques with robust classifiers to improve predictive accuracy in educational settings, particularly for imbalanced [16]. Another study on educational data used the High School Longitudinal Study of 2009 (HSLS: 09) dataset to classify students into those likely to pursue higher education and those expected to defer or drop out. Sampling methods like ROS, RUS, SMOTE-NC, and hybrid techniques were tested, with hybrid resampling performing best for highly imbalanced datasets when paired with random forest classifiers [17]. Lastly, a study on physical fitness data classification analyzed relationships in students' physical fitness using a dataset of 812 records. The data, which included six fitness-related attributes such as BMI and exercise performance metrics, was classified using decision trees, random forests, and association rule mining techniques. The decision tree technique achieved a maximum accuracy of 100%, while random forests reached 99.8%. Additionally, association rule mining via Apriority and FP-Growth produced consistent patterns in the data. Collectively, these studies underscore the importance of advanced data balancing methods like SMOTE and hybrid resampling, as well as robust machine learning models such as random forests, in effectively addressing class imbalances across various datasets and applications. Tools like WEKA and Python were pivotal in facilitating these analyses, enabling efficient preprocessing, data balancing, and model evaluation [18]. Collectively, these studies highlight the effectiveness of data balancing techniques, particularly SMOTE and hybrid methods, and machine learning models like random forests in improving classification performance. These findings emphasize the critical role of advanced sampling methods and robust algorithms in addressing imbalanced data challenges across various domains, enabling more accurate predictions and better insights into complex datasets.

III. Methodology

A. Dataset

This section outlines the procedures used for data collection, which were carefully designed to ensure accuracy, consistency, and participant safety. The key steps involved are as follows:

Participants:

The study involved 75 male football student-athletes from Nakhon Phanom Sports School, Nakhon Phanom Province, Thailand, aged between 11 and 16 years. The participants were selected based on the following criteria:

• Being enrolled in the youth football training program under the school's athletic development system

• Being in good physical health with no injuries at the time of testing

• Having received informed consent from a parent, guardian, or coach

Data Collection Procedure:

1) *Preparation Phase*

Participants were informed in advance about the testing procedures. All athletes underwent a physical condition check and were instructed to perform warm-up exercises before each test. Testing environments were arranged according to the requirements of each physical test to ensure safety and measurement accuracy.

2) *Testing Phase*

All physical fitness tests were conducted individually. Each measurement was administered by trained evaluators or members of the research team. Data from each test was recorded immediately using standardized forms and later entered into a computer system for further analysis. The tests are divided into four aspects, which include:

Flexibility: Measured using the Sit and Reach test, with measurements recorded in centimeters to evaluate forward trunk flexibility.

Leg Muscle Strength: Measured using a Leg Strength Dynamometer, with results expressed in kilograms per body weight (kg/BW) to account for individual weight differences and provide a normalized measure of lower limb strength.

Muscular Endurance and Strength: Assessed using the 1-Minute Sit-Up test, with the number of correctly completed sit-ups recorded within one minute. The results were measured in repetitions.

Body Fat: Measured using a Body Composition Analyzer (The ACCUNIQ BC-360). The results were expressed as a percentage (%) of total body mass.

3) *Data Cleansing*

Collected data were verified for accuracy and consistency. Outliers were identified by comparing values against standard physical fitness benchmarks set by the Sports Authority of Thailand, Region 3 [19]. In cases where anomalies or measurement errors were detected, retesting was conducted for the concerned athlete, if necessary.

The physical fitness test results (Fitness Level) could be categorized into five levels: Very low (poor), low (fair), average, good, and very good (excellent). An example of the athletes' physical fitness test data is presented in Table I, while Table II summarizes the basic statistical values across the four main components of physical fitness.

TABLE I
EXAMPLES OF PHYSICAL FITNESS TEST DATA

No.	Sit and Reach (cm)	Leg Strength (kg/BW)	Sit Up 1 min (number)	Body Fat (%)
1	10.00	3.34	45	14.20
2	12.00	2.42	43	18.50
3	17.00	2.62	47	10.90
4	15.00	2.69	49	13.50
5	10.00	2.95	35	19.00

TABLE II
DESCRIPTIVE STATISTICS FOR PHYSICAL FITNESS TEST RESULTS

Finess Test	Min	Max	Average	Deviation
Sit and Reach (cm)	3.00	28.00	15.60	5.51
Leg Strength (kg/BW)	1.38	4.54	2.48	0.61
Sit Up 1 min. (number)	18.00	62.00	43.69	7.17
Body Fat (%)	3.00	28.60	14.49	5.77

From the analysis and evaluation of the physical fitness of 75 participants, it was found that 5 athletes were in the excellent category, 17 athletes were in the Good category, 23 athletes were in the Average category, 22 athletes were in the Fair category, and 8 athletes were in the Poor category. These figures represent 6.70%, 22.70%, 30.70%, 29.20%, and 10.70% of the total population, respectively. A bar chart showing the number of participants categorized by physical fitness test results is presented in Fig. 1.

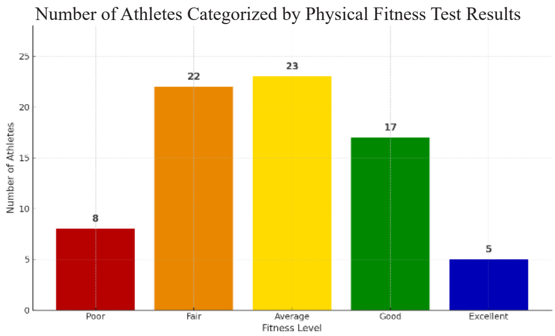

Fig. 1. Distribution of athletes by fitness level

It can be observed that the sample size is relatively small, and the proportion of data in each class varies significantly. The approach to addressing the issue of class imbalance will be discussed in the next section.

B. The Application of SMOTE to Handling Imbalanced Data

Several sampling methodologies have been developed to address class imbalance in supervised learning, including both under-sampling and oversampling techniques. Under sampling methods, such as Random Under sampling and Tomek Links, reduce the majority class size to balance the dataset, but risk losing valuable information. Oversampling approaches, on the other hand, expand the minority class by duplicating existing data or generating synthetic samples. Among these, SMOTE (Synthetic Minority Oversampling Technique) and ADASYN (Adaptive Synthetic Sampling) are widely used. While both generate synthetic samples, SMOTE creates them uniformly between minority class neighbors, whereas ADASYN focuses more on difficult-to-learn areas near the decision boundary. In this study, SMOTE was preferred over ADASYN due to its stability and balanced sample generation, which are especially important when working with small datasets. ADASYN, while effective in complex scenarios, can amplify noise and increase the risk of overfitting in limited or noisy datasets. Therefore, SMOTE was selected to ensure more controlled and interpretable oversampling outcomes [13], [20]. A total of 86 new samples were synthesized, resulting in a total of 161 records. These were classified as follows: 30 records for athletes with excellent performance, 32 records with good performance, 34 records with average performance, 33 records with fair performance, and 32 records with poor performance. This corresponds to proportions of 18.60%, 19.90%, 21.10%, 20.50%, and 19.90%, respectively. Table III shows the actual data and synthetic data for each class. Fig. 2 presents a bar chart showing actual data together with synthetic data classified by class. Section 4 presents and compares the experimental results.

TABLE III
THE NUMBER OF THE ACTUAL DATA AND SYNTHETIC DATA IN EACH CLASS

Class	Actual data	Synthetic data	Total
Poor	8	24	32
Fair	22	11	33
Average	23	11	34
Good	17	15	32
Excellent	5	25	30
TOTAL	**75**	**86**	**161**

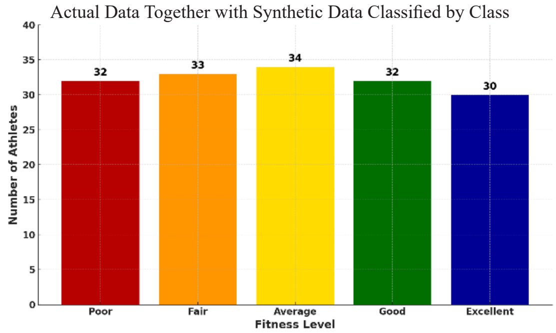

Fig. 2. Actual and synthetic data classified by class

It can be observed that this study synthesized a greater number of new data records compared to the original dataset. From the review of related studies, it was found that several studies have successfully applied the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic samples, often exceeding the number of original minority class samples, to address class imbalance and enhance model performance. Notable examples include: [13], [21]-[23]. These studies collectively demonstrate the successful application of SMOTE and its variants in generating synthetic samples that exceed the original number of minority class samples, leading to improved classifier performance in imbalanced datasets.

C. Tools used for Data Analysis

This study utilized RapidMiner Studio 9.10 [24] in conjunction with the Weka Extension [25] and the Python Scripting Extension [26] for both data preparation and developing models to evaluate the physical fitness of athletes. RapidMiner Studio is an integrated software designed to provide convenience and efficiency for data science tasks such as data mining and machine learning. It offers a wide range of operators, categorized by problem type, allowing users to select the most appropriate solution for their tasks. It could also be seamlessly integrated with other tools such as Weka Extension, Python Script, and Jupyter Notebook.

D. The Supervised Learning Model for Evaluating the Physical Fitness of Athletes

This study employed five supervised learning models, including Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM (Python Learner). These models were implemented to evaluate datasets effectively, offering unique advantages and trade-offs. The brief details and working principles of these models are as follows [24]-[27]:

1) Decision Tree

The Decision Tree model is a simple yet powerful algorithm for classification and regression tasks. It works by splitting the dataset into subsets based on feature values, forming a tree-like structure. Each node represents a decision rule, and leaf nodes represent outcomes. Decision Trees are easy to interpret and visualize, suitable for datasets with mixed data types, but are prone to overfitting, especially with noisy data, unless pruned effectively.

2) Random Forest

Random Forest is an ensemble learning method that builds multiple decision trees and combines their outputs to improve accuracy and robustness. It addresses the overfitting issue of single decision trees by averaging predictions. Random Forests offer high accuracy and robustness to noise, are effective for handling large datasets and high-dimensional data, but are computationally intensive compared to single decision trees.

3) Neural Network

Neural Networks mimic the structure and functionality of the human brain, comprising layers of interconnected neurons. This model excels at identifying complex patterns and relationships in data. Neural Networks are versatile in handling structured and unstructured data, deliver high performance for large datasets with sufficient computational power, but require careful tuning of hyper parameters and may be prone to overfitting without sufficient training data.

4) Multinomial Logistic Regression

Multinomial Logistic Regression is a statistical model used for multi-class classification problems. It extends logistic regression by modeling the probability of each class as a function of input features. This model is simple and easy to implement, performs well with linearly separable data, but has limited ability to capture non-linear relationships in complex datasets.

5) LightGBM (Python Learner)

Python Learner in RapidMiner allows integration with Python scripts, offering flexibility for implementing custom algorithms. LightGBM (Light Gradient Boosting Machine), a widely used model implemented using Python Learner, is a high-performance gradient boosting framework based on decision tree algorithms. This model is fast, scalable, and efficient for large datasets, supports both continuous and categorical features, and is particularly effective for structured data. Unlike simpler models, LightGBM captures complex feature interactions and does not assume feature independence, making it well-suited for a wide range of real-world classification and regression tasks.

E. Model Performance Evaluation

This study utilizes the Confusion Matrix, a widely used tool for evaluating the outcomes of predictions or estimations made by developed models [27]. An example of a 2x2 Confusion Matrix is illustrated in Fig. 3.

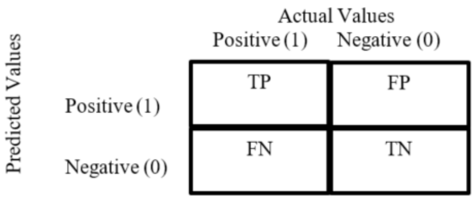

Fig. 3. Confusion matrix structure

Where:

TP: True Positive
FN: False Negative
FP: False Positive
TN: True Negative

The accuracy can be calculated by taking the percentage of correct predictions out of the total number of samples. Correct predictions are defined as instances where the predicted attribute value matches the actual target attribute value. The accuracy can be calculated as follows:

$$Accuracy = \left(\frac{Correct\ Predictions}{Total\ Samples}\right) * 100 \quad (2)$$

IV. EXPERIMENTAL RESULTS

For training and testing the model, the physical fitness test data was divided into two subsets: 70% was used to train the models, while the remaining 30% was reserved for evaluating model performance. Although the data was initially shuffled, stratified sampling was applied to ensure that the class distribution in both the training and testing sets reflected that of the overall dataset. It is important to note that the performance evaluation in this study was conducted solely on the test dataset.

This section presents the assessment results of athletes' physical fitness using both the original dataset and the newly synthesized dataset.

A. The Assessment Results of the Original Dataset

The assessment results of physical fitness assessment using Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM techniques to classify the original dataset consisting of 75 records, as shown in Tables IV to VIII, respectively.

TABLE IV
THE ASSESSMENT RESULTS USING THE DECISION TREE: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	0	0	0	0	0.00%
pred. Good	1	3	0	0	0	75.00%
pred. Average	0	1	5	0	0	83.33%
pred. Fair	0	1	2	6	0	66.67%
pred. Poor	0	0	0	1	2	66.67%
class recall	0.00%	0.60%	71.43%	85.71%	100.00%	

Accuracy: 72.73%

TABLE V
THE ASSESSMENT RESULTS USING THE RANDOM FOREST: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	0	0	0	0	0.00%
pred. Good	1	2	0	0	0	66.67%
pred. Average	0	3	5	1	0	55.56%
pred. Fair	0	0	2	5	1	62.50%
pred. Poor	0	0	0	1	1	50.00%
class recall	0.00%	40.00%	71.43%	71.43%	50.00%	

Accuracy: 59.09%

TABLE VI
THE ASSESSMENT RESULTS USING THE NEURAL NETWORK: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	1	1	0	0	0	50.00%
pred. Good	0	3	0	0	0	100.00%
pred. Average	0	1	5	0	0	83.33%
pred. Fair	0	0	2	6	0	75.00%
pred. Poor	0	0	0	1	2	66.67%
class recall	100.00%	60.00%	71.43%	85.71%	100.00%	

Accuracy: 77.27%

TABLE VII
THE ASSESSMENT RESULTS USING THE MULTINOMIAL LOGISTIC REGRESSION: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	1	1	0	0	0.00%
pred. Good	1	3	0	0	0	75.00%
pred. Average	0	1	5	0	0	83.33%
pred. Fair	0	0	1	6	0	85.71%
pred. Poor	0	0	0	1	2	66.67%
class recall	0.00%	60.00%	71.43%	85.71%	100.00%	

Accuracy: 72.73%

TABLE VIII
THE ASSESSMENT RESULTS USING THE LightGBM: THE ORIGINAL DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	0	0	1	0	0	0.00%
pred. Good	0	3	1	0	0	75.00%
pred. Average	1	2	4	0	0	57.14%
pred. Fair	0	0	1	6	2	66.67%
pred. Poor	0	0	0	1	0	0.00%
class recall	0.00%	60.00%	57.14%	85.71%	0.00%	

Accuracy: 59.09%

B. *The Assessment Results of the New Dataset*

The Confusion Matrix tables present the assessment results of physical fitness assessment using Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM techniques to classify the original dataset combined with synthetic data, 161 records in total shown in Tables IX-XIII, respectively.

TABLE IX
THE ASSESSMENT RESULTS USING THE DECISION TREE: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	6	2	1	0	0	66.67%
pred. Good	3	8	0	0	0	72.73%
pred. Average	0	0	6	1	0	85.71%
pred. Fair	0	0	3	9	0	75.00%
pred. Poor	0	0	0	0	10	100.00%
class recall	66.67%	80.00%	60.00%	90.00%	100.00%	

Accuracy: 79.59%

TABLE X
THE ASSESSMENT RESULTS USING THE RANDOM FOREST: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	6	2	0	0	0	75.00%
pred. Good	1	8	0	0	0	88.89%
pred. Average	0	0	7	2	0	77.78%
pred. Fair	2	0	3	8	0	61.54%
pred. Poor	0	0	0	0	10	100.00%
class recall	66.67%	80.00%	70.00%	80.00%	100.00%	

Accuracy: 79.59%

TABLE XI
THE ASSESSMENT RESULTS USING THE NEURAL NETWORK: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	8	2	0	0	0	80.00%
pred. Good	1	7	2	0	0	70.00%
pred. Average	0	1	7	1	0	77.78%
pred. Fair	0	0	1	9	2	75.00%
pred. Poor	0	0	0	0	8	100.00%
class recall	88.89%	70.00%	70.00%	90.00%	80.00%	

Accuracy: 79.59%

TABLE XII
THE ASSESSMENT RESULTS USING THE MULTINOMIAL LOGISTIC REGRESSION: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	7	3	0	0	0	70.00%
pred. Good	2	6	0	0	0	75.00%
pred. Average	0	1	9	0	0	90.00%
pred. Fair	0	0	1	8	3	66.67%
pred. Poor	0	0	0	2	7	77.78%
class recall	77.78%	60.00%	90.00%	80.00%	70.00%	

Accuracy: 75.51%

TABLE XIII
THE ASSESSMENT RESULTS USING THE LightGBM: THE NEW DATASET

	true Exellent	true Good	true Average	true Fair	true Poor	class precision
pred. Exellent	9	1	0	0	0	90.00%
pred. Good	0	8	0	0	0	100.00%
pred. Average	0	1	9	0	0	81.82%
pred. Fair	0	0	1	2	2	75.00%
pred. Poor	0	0	0	8	8	100.00%
class recall	100.00%	80.00%	90.00%	90.00%	80.00%	

Accuracy: 87.76%

C. *Comparison of Results*

The assessment accuracy of all five models on the two physical fitness datasets is presented in Table XIV.

TABLE XIV
A SUMMARY OF THE ASSESSMENT ACCURACY OF ALL FIVE MODELS ON FIVE MODELS ON THE TWO PHYSICAL FITNESS DATASETS

List	The Original Dataset (%)	The New Dataset (%)
Decision Tree	72.73	79.59
Random Forest	59.09	79.59
Neural Network	77.27	79.59
Multinomial Logistic Regression	72.73	75.51
LightGBM	59.09	87.76
Average	68.18	80.40

The table demonstrates that in the evaluation of physical fitness using the original dataset, the Neural Network model performed the best, achieving the highest accuracy of 77.27%. It was followed by the Decision Tree and Multinomial Logistic Regression models, both with an accuracy of 72.73%. The Decision Tree and LightGBM models showed the lowest accuracy at 59.09%. For the dataset consisting of 161 records, original data combined with 86 synthetic records generated using the SMOTE method, the LightGBM (Python Learner) model achieved the highest accuracy of 87.76%. The Decision Tree, Random Forest, and Neural Network models followed, with the highest accuracy being 79.59%. The Multinomial Logistic Regression model showed the lowest accuracy in evaluating this dataset, with an accuracy of 75.51%. As observed, the evaluation of physical fitness using the original dataset combined with the newly synthesized data, totaling 161 records, showed improved performance. The highest accuracy reached 87.76%, achieved by the LightGBM (Python Learner) model, which had an average accuracy of 81.41%. This represents an increase of 12.23% compared to the average performance of the evaluation using only the original dataset. The analysis revealed that LightGBM outperformed other supervised learning models due to its highly efficient gradient boosting framework. Unlike traditional tree construction methods, LightGBM builds trees in a leaf-wise manner, resulting in greater loss reduction and improved accuracy. Compared to other models such as logistic regression, decision trees, or Support Vector Machines (SVMs), LightGBM handles large-scale and high-dimensional datasets with superior speed and precision, thanks to its histogram-based computation and advanced sampling strategies. Moreover, it natively supports missing values and categorical features, reducing the need for extensive pre-processing. When implemented through the Python Learner in RapidMiner Studio, LightGBM offers flexible hyper parameter tuning and integrates seamlessly into the analytical workflow, making it a powerful and scalable solution for predictive modelling tasks. While synthetic data generation techniques such as SMOTE can effectively address class imbalance and improve model performance, they may also introduce certain biases. For instance, synthetic samples are created by interpolating between existing minority class instances, which can lead to the over-representation of specific regions in the feature space while neglecting others. This may result in models that generalize poorly or are overly confident in areas where no real data exists. Additionally, if the original dataset contains noise or mislabelled instances, synthetic sampling may inadvertently amplify these issues. Therefore, careful validation and data quality checks are essential when using synthetic data to ensure that it enhances rather than distorts the learning process.

V. Conclusions

This research collected physical fitness test data from 75 student-athletes aged 11-16 at the Nakhon Phanom Sports School. The tests were divided into four aspects: flexibility, leg muscle strength, endurance and muscle strength, and body fat. The collected data underwent cleaning (Data Cleansing) to ensure accuracy and was initially evaluated using the standard youth athlete fitness test criteria of the Sports Authority of Thailand, Region 3. The physical fitness test results (Fitness Level) were classified into five levels: Poor, Fair, Average, Good, and Excellent. Analysis revealed that 5 students were in the Excellent category, 17 in Good, 23 in Average, 22 in Fair, and 8 in Poor, accounting for 6.70%, 22.70%, 30.70%, 29.20%, and 10.70% of the total population, respectively. This dataset is imbalanced, which could affect the performance of models used for data processing. To address this imbalance and improve supervised learning model performance, this study utilized the Synthetic Minority Oversampling Technique (SMOTE), a special oversampling method that generates new synthetic data points instead of replicating existing ones. The machine learning models used for evaluating the student-athlete dataset included Decision Tree, Random Forest, Neural Network, Multinomial Logistic Regression, and LightGBM (Python Learner). The findings revealed that applying SMOTE to increase the dataset size to 161 records with SMOTE improved performance, achieving a maximum accuracy of 87.76%. The LightGBM (Python Learner) model demonstrated an average accuracy of 81.41%, representing a 12.23% improvement compared to the average performance with the original dataset. In summary, the results indicate that using oversampling techniques like SMOTE can mitigate data imbalance issues and enhance data classification or evaluation performance when the additional data is generated in an appropriate quantity. However, other factors may also influence model performance and data classification efficiency. In summary, these models, such as LightGBM and other supervised learning techniques, could be implemented in real-world athlete training programs to monitor performance, predict injury risks, and personalize training plans based on physiological data. By leveraging predictive analytics, coaches and sports scientists can make data-driven decisions that optimize performance and reduce overtraining. Future research could explore the integration of real-time data from wearable devices and expand predictive modelling to include psychological, nutritional, and environmental factors for a more holistic view of athlete development.

Acknowledgment

We would like to express our gratitude to the Nakhon Phanom Sports School for their support in facilitating the physical fitness testing of student-athletes. Additionally, we extend our sincere thanks to the Faculty of Management Sciences and Information Technology, Nakhon Phanom University, for providing funding support for this study.

References

[1] R. Sharda, D. Delen, and E. Turban, *Business Intelligence, Analytics, and Data Science: A Managerial Perspective*, 4th ed. Harlow, UK: Pearson, 2018, pp. 33-36.

[2] R. C. Prati, G. E. Balista, and M. C. Monard, “Data mining with imbalanced class distributions: concepts and methods,” in *Proc. Indian Int. Conf. Artif. Intell., IICAI 2009*, 2009, pp. 359-376.

[3] G. Lemaître, F. Nogueira, and C. Aridas, “Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning,” *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1-15, Sep. 2016.

[4] U. Ninrutsirikun, H. Imai, B. Watanapa, and C. Arpnikanondt, “Principal component clustered factors for determining study performance in computer programming class,” *Wireless Pers. Commun.*, vol. 115, no. 4, pp. 2897-2916, Dec. 2020.

[5] R. Liu, “A novel synthetics minority oversampling technique based on relative and absolute densities for imbalanced classification,” *Appl. Intell.*, vol. 53, pp. 786-803, Apr. 2003.

[6] H. He and E. A. Garcia, “Learning from imbalanced data,” *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263-1284, Sep. 2009.

[7] Q. Yang and X. Wu, “10 challenging problems in data mining research,” *Int. J. Inf. Technol. Decis. Mak.*, vol. 5, no. 4, pp. 597-604, Dec. 2006, https://doi.org/10.1142/S0219622006002258

[8] F. Pedregosa et al., “Scikit-learn: Machine learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825-2830, Oct. 2011.

[9] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, 2016, pp. 85-794.
[10] A. Vaswani et al., “Attention is all you need,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 5998-6008, Aug. 2017, https://doi.org/10.48550/arXiv.1706.03762
[11] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Adv. Neural Inf. Process. Syst.*, vol. 30, pp. 4765-4774, Nov. 2017, https://doi.org/10.48550/arXiv.1705.07874
[12] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770-778.
[13] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321-357, Jun. 2002, https://doi.org/10.1613/jair.953
[14] P. Thanathammathee and Y. Sirisathitkul, “Improved classification techniques for imbalanced dataset of elderly’s knee osteoarthritis,” *J. Sci. Technol.*, vol. 27, no. 6, pp. 1164-1178, Nov.-Dec. 2019.
[15] S. Wannont and R. Muangsarn, “Improving prediction models of student business career using sampling techniques for learning in multi-classes imbalance dataset,” *Chaiyaphum Parithat J.*, vol. 4, no. 1, pp. 39-49, Jan.-Apr. 2021.
[16] N. Rachburee and W. Punlumjeak, “Oversampling technique in student performance classification from engineering course,” *Int. J. Electr. Comput. Eng.*, vol. 11, no. 4, pp. 3567-3574, Aug. 2021.
[17] T. Wongvorachan, S. He, and O. Bulut, “A comparison of undersampling, oversampling, and SMOTE methods for dealing with imbalanced classification in educational data mining,” *Information*, vol. 14, no. 1, p. 54, Jan. 2023.
[18] C. Masachai, N. Srisahno, W. Masachai, and W. Buathong, “Relationship physical fitness assessment results of students with data mining at Rajaprachanugroh 1 school,” *Ind. Technol. Lampang Rajabhat Univ.*, vol. 14, no. 2, pp. 1-11, Jul.-Dec. 2021.
[19] S. Kusum, C. Chiewsakul, J. Naksri, N. Mudchanthuek, and W. Deeniwong, “Physical fitness test and standard guidelines for youth athletes,” *Regional Sports Sci. Work, Sports Authority of Thailand, Region 3, Ministry of Tourism and Sports*, 2019. [Online]. Availabe: https://set3.org [Accesied: Apr. 25, 2024]
[20] H. He, Y. Bai, E. A. Garcia, and S. Li, “ADASYN: Adaptive synthetic sampling approach for imbalanced learning,” in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN)*, 2008, pp. 1322-1328.
[21] D. Elreedy and A. F. Atiya, “A theoretical distribution analysis of the synthetic minority oversampling technique,” *Mach. Learn.*, vol. 111, no. 1, pp. 157-180, Jan. 2024.
[22] R. Liu, “A novel synthetic minority oversampling technique based on relative and absolute densities for imbalanced classification,” *Appl Intell*, vol. 53, p. 786803, Jan. 2023, https://doi.org/10.1007/s10489-022-03512-5
[23] G. Douzas and F. Bacao, “Geometric SMOTE: Effective oversampling for imbalanced learning through a geometric extension of SMOTE,” *Inf. Sci.*, vol. 465, pp. 1-20, Sep. 2017.
[24] RapidMiner, Inc., “RapidMiner Studio, version 9.10.” *Docs. Rapidminer. Com*. 2024. [Online]. Available: https://www.rapidminer.com [Accessed May 25, 2024].
[25] M. Hall, E. Frank, and G. Holmes, “The WEKA data mining software: An update,” *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10-18, Nov. 2009, https://doi.org/10.1145/1656274.1656278
[26] G. van Rossum and F. L. Drake, “Python 3 reference manual,” *(S. l.) ACM DL.* 2009. [Online]. Available: https://dl.acm.org/doi/abs/10.5555/1593511 [Accessed Mar. 20, 2009].
[27] R. Kohavi and F. Provost, “Glossary of terms,” *Mach. Learn.*, vol. 30, no. 2-3, pp. 271-274, Jan. 1998.

Janyarat Phrueksanant received her M.Sc. in Information Technology from the School of Information Technology, King Mongkut’s University of Technology Thonburi, and Ph. D. in Systems Engineering from the School of Engineering, Cardiff University. She currently works as a lecturer at the Department of Information Technology and Computer Innovation, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Chayanont Awikunprasert received his M.Sc. in Human Development from the Faculty of Graduate Studies, Mahidol University, and Ph.D. in Exercise and Sport Science from the Faculty of Sport Science, Burapha University. He currently works as a lecturer at the Department of Sports Science, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Jirachai Karawa received his B.Sc. in Sport Science from the Faculty of Education, Mahasarakham University, and M.Sc. in Exercise and Sport Sciences from the Graduate School, Khon Kaen University. He currently works as a lecturer at the Department of Sports Science, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Sutthirak Wisetsang received his B.Ed. and M.Ed. in Physical Education from the Faculty of Education, Ramkhamhaeng University. He currently works as a lecturer at the Department of Sports Science, Faculty of Management Sciences and Information Technology, Nakhon Phanom University.

Predictive Analysis of Academic Achievement in Information Studies: A Comparative Study Using Educational Data Mining Techniques

Knitchepon Chotchantarakun*
Department of Information Studies, Faculty of Humanities and Social Sciences,
Burapha University, Chonburi, Thailand
E-mail: knitchepon@go.buu.ac.th*

Received: March 19, 2025 / Revised: May 29, 2025 / Accepted: June 5, 2025

***Abstract*—Predicting students' academic achievement in the initial stages is beneficial for designing effective training programs to enhance success rates. Extracting knowledge from student data is a fundamental aspect of Educational Data Mining (EDM). This study aims to analyze the predictive factors influencing the outcomes of graduates from the Information Studies program. The results not only contribute to improving student performance but also aid in constructing a better curriculum. A dataset is utilized within five classification models to categorize students into four target classes. The datasets are grouped into three types: Demographic information, course grades, and early-stage GPA. This study addresses the issue of the imbalanced dataset by applying the Synthetic Minority Over-sampling Technique (SMOTE). The findings indicate that early-stage GPA (90.5%) is the most significant predictor, particularly when applying the Naive Bayes classifier on a balanced dataset. In contrast, demographic information (58.0%) and core course grades (87.5%) show lower predictive influence. The findings support learning strategies and enhancing curriculum design to improve final academic outcomes.**

***Index Terms*—Classification, Educational Data Mining, Feature Selection, Imbalanced Dataset, Machine Learning**

I. Introduction

With the transition from paper-based documentation to digital formats, educational institutions generate vast volumes of electronic data. Transforming this extensive data into meaningful knowledge is essential in decision-making, improving learning quality, and providing information for institutional planning to maximize efficiency. The extraction of such knowledge has been facilitated by advancements in computer technology, particularly in Artificial Intelligence (AI), Data Mining (DM), and Machine Learning (ML). These progressions enable the application of predictive modeling, clustering, and association rules mining to identify patterns and correlations within educational data. These techniques contribute to the refinement of academic curricula, enhancement of teaching methodologies, and improvement of student learning outcomes.

The DM process encompasses a wide range of techniques and algorithms that are applied in various domains, including medicine, marketing, industry, finance, and education. The specific application of data mining to educational data is referred to as Educational Data Mining (EDM) [1]. Baker and Yacef identified five primary approaches within EDM: Prediction, clustering, relationship mining, distillation of data for human judgment, and discovery with models. Recent works usually integrate a combination of prediction, clustering, and data distillation for human judgment.

EDM has emerged as a significant area of research, driven by advancements in educational database management systems. It emphasizes on developing specialized and quantitative methodologies to analyze large volumes of data collected from institutions across various educational levels. By applying DM techniques to student datasets, educators can gain deeper insights into student behavior and performance. While academic achievement is traditionally assessed through metrics such as course grades and Grade Point Average (GPA), demographic factors also play a crucial role in shaping educational outcomes. Consequently, the application of DM to academic and demographic data enhances the program design at the department, faculty, and institution levels.

Student academic success is a key indicator of the teaching and learning quality of the institutions. Early classification of students during the initial stages of their study is an effective approach to minimizing the possibility of dropout. Additionally, this strategy contributes to improved academic performance by facilitating optimal resource allocation within institutions, ensuring that resources are utilized effectively. Furthermore, it supports the development of potential students, enabling them to gain higher academic achievements.

Although data obtained from the university's Student Information System (SIS) is increasingly utilized in higher education, many institutions still face challenges in developing effective predictive models. Normally, static indicators such as course grades or test scores are used to estimate the student's learning outcome. However, they cannot reflect the complexity and dynamics of learning behavior influencing academic success. As a result, opportunities for the timely identification and support of low-performance students are often missed. There is a critical need for predictive models that are both data-driven and interpretable, particularly those that leverage early academic indicators such as demographic information, course grades, and GPA. These models must provide prominent levels of predictive accuracy and be meaningful enough to enable educators to implement early support and enhance instructional strategies.

This study investigates the application of machine learning techniques to identify the key factors influencing students' final academic achievement during the initial stages of their educational journey. The prediction is conducted using five classification models within the EDM framework. Experiments are performed using undergraduate students' data from the Information Studies (IS) program, combined with the Synthetic Minority Over-sampling Technique (SMOTE) to address the imbalanced dataset. This research contributes to enhancing students' final academic outcomes and assisting with the curriculum design.

The paper is organized as follows: Section II reviews relevant literature. Section III outlines the research methodology employed in this study. Section IV presents the experimental results along with explanatory analysis. Section V offers a detailed discussion of the findings. Finally, Section VI concludes the paper with a summary of key findings and significant issues contributing to the learning strategies.

II. Literature Review

Early prediction of student outcomes enables educators and administrators to make informed, timely decisions to enhance course effectiveness. It facilitates the development of specialized training programs aimed at increasing student success rates. Advancements in EDM have shown the application of DM techniques across various educational domains [2]. Research objectives in this field can be defined at multiple levels, including degree, academic year, course, and examination levels. Studies in this area commonly employ classification techniques, such as predicting student outcomes as Pass or Fail, as well as regression techniques, such as estimating the Cumulative Grade Point Average (CGPA). Typically, CGPA prediction utilizes students' Grade Point Averages (GPA) from their first two years to forecast their final CGPA at graduation.

EDM plays a crucial role in uncovering patterns and insights related to educational phenomena and learning processes [5], as well as in understanding students' academic performance. EDM has been widely applied to predict academic outcomes across various domains, including academic performance [4], student retention [5], study success [6], academic satisfaction [7], and dropout rates [8].

Key factors influencing education, commonly explored in EDM, include prior academic achievement, student demographic characteristics, e-learning activities, psychological aspects, and the learning environment. A study [9] highlighted that 69% of research in this field utilizes pre-university academic performance and demographic characteristics of learners. Among the most frequently used predictors of academic achievement are student assessments and cumulative Grade Point Averages (GPAs). Both pre-university information and data collected during the study period, such as semester grades and GPAX, significantly influence the prediction of academic achievement [10]. These elements, derived from students' academic journeys, are critical in forecasting their overall academic success throughout their educational tenure.

In this study [11], DM techniques were employed to analyze the academic achievement of 210 undergraduate students. The authors developed a predictive model to estimate students' final academic performance and explored the relationship between their academic outcomes and progress during their course of study. The variables utilized in the analysis were exclusively related to scores or grades. A decision tree algorithm was applied to construct a classification model, which categorized the dataset based on four information criteria: Information Gain, Gini Index, Accuracy, and Gain Ratio. Additionally, the X-means algorithm, using Euclidean distance and the Bayesian Information Criterion (BIC), was employed to group students into categories reflecting high and low academic performance. The DM process was implemented using RapidMiner. The findings from this predictive model offer early intervention opportunities for students identified in the low-performing groups while also providing guidance and opportunities for those demonstrating strong academic performance.

In 2022, a study [12] conducted a comprehensive review and analysis of emerging literature on the application of Artificial Neural Network (ANN) to predict academic achievement among university students. The article highlighted that ANN techniques are frequently combined with other data mining methods to identify patterns and assess academic performance. EDM research focuses on the university level, as most researchers are affiliated with universities and have easier access to student data. Furthermore, ANN often demonstrates higher accuracy in evaluating model performance compared to other algorithms. CGPA was identified as a commonly used predictive factor, while other factors were found to have a minimal impact on the model's performance.

Reference [13] conducted a study on the academic performance of 635 master's degree candidates across diverse faculties, such as Business Administration, Engineering, and Information Technology. The study employed six machine learning algorithms, utilizing CGPA as the principal predictor. Evaluation of model efficacy utilized Mean Square Error (MSE) and Mean Absolute Error (MAE) metrics to gauge the disparity between predicted and actual GPA scores. Findings indicated that the Neural Network (NN) algorithm exhibited superior performance, attaining the lowest error rates and demonstrating heightened predictive precision compared to other models.

A comprehensive survey and synthesis of 402 articles related to EDM and Learning Analytics (LA) were examined in a review study [14]. The study highlighted that EDM and LA techniques are effective in addressing various learning-related challenges. The application of DM techniques across these studies included Classification (26.23%), Clustering (21.25%), Image Mining (15%), Statistical Analysis (14.25%), Association Rule Mining (14%), Regression (10.25%), and Sequential Pattern Mining (6.5%), etc. These techniques contributed to the development of improved learning strategies for students. Additionally, the review introduced researchers to appropriate methodologies for conducting research in different educational contexts and provided essential tools and information to enhance university education systems. Reference [15] evaluated the effectiveness of ML techniques in predicting student performance using various ML algorithms. The research considered factors such as data quality, feature selection, and model complexity. The findings demonstrated that certain ML methods are particularly effective in forecasting student performance, thereby offering valuable insights for decision-making and academic planning.

Previous research has demonstrated that predicting student academic achievement can achieve high levels of accuracy, particularly when EDM utilizes classification techniques with a limited number of categories. For example, predicting binary outcomes such as pass/fail status or categorizing students based on satisfactory performance often results in even greater predictive accuracy. However, the effectiveness of these predictive models is influenced by several factors, including the choice of algorithms, the selection of variables, and the size of the dataset. In addition, demographic attributes such as age, gender, religion, place of residence, family background, employment status, and past GPA have been identified as significant contributing factors that enhance the accuracy of academic achievement predictions [16], [17].

Regarding the application of SMOTE, this research [18] demonstrated that applying SMOTE before splitting the dataset into training and testing sets enhanced the accuracy of the ANN model. Specifically, it yielded accuracy improvements ranging from 1.94% to 3.98% across multiple datasets, highlighting its capability to address class imbalance and improve the overall performance of classification models. This study [19] presented the improvement in the accuracy of ANN on imbalanced datasets using SMOTE by generating synthetic samples to balance the class distribution. However, this process may introduce noise into the dataset. The proposed method addresses this limitation by incorporating an Autoencoder, which helps filter out noise and enhances the overall classification performance.

Despite the growing impact of EDM in improving student outcomes across higher education, current research disproportionately focuses on STEM (Science, Technology, Engineering, and Mathematics) disciplines and general education courses. In contrast, specialized academic domains such as the IS program remain significantly underrepresented, which presents a critical gap in the literature. This program emphasizes interdisciplinary knowledge, including critical thinking, system analysis, information behavior, digital literacy, and information retrieval. These competencies are typically assessed through project work and collaborative assignments rather than focusing only on the numerical scores and exam-based evaluations commonly used in STEM. As a result, existing EDM models may not capture the detailed indicators of academic success relevant to information studies students.

Additionally, the demographic and academic profiles of students in this program may differ from those in traditional STEM fields due to diverse academic backgrounds. This research explores this gap by applying EDM methods to analyze academic performance within the IS program. The findings are expected to contribute to the growing body of EDM literature while generating actionable insights for improving teaching, learning, and student support in the field related to information science. Consequently, the use of predictive approaches on these types of datasets is still limited, indicating a clear research gap. This study aims to address this underexplored area by analyzing this specific population, which may yield novel insights to inform curriculum design and enhance student support within these academic domains.

III. Research Methodology

A. Educational Data Mining Process

The EDM process [20] comprises six key stages as shown in Fig. 1. The data collection phase involves gathering information from multiple sources, including pre-enrollment records, demographic details, students' learning environments, academic performance, and psychological attributes. These data are primarily obtained from the university's Student Information System (SIS), supplemented by student surveys. Once collected, the data are prepared for subsequent analysis.

During the initial preparation stage, the raw data are transformed into a structured format through a series of processes, including 1) Selection, 2) Cleaning, and 3) Derivation of new variables. This phase is particularly critical and often requires the most time to complete. Following this, statistical analysis serves as an initial exploratory step, offering an overview of the dataset and assisting researchers in understanding key characteristics before proceeding with data mining. This analysis typically includes descriptive statistics such as frequency, mode, median, mean, standard deviation, variance, range, and correlation.

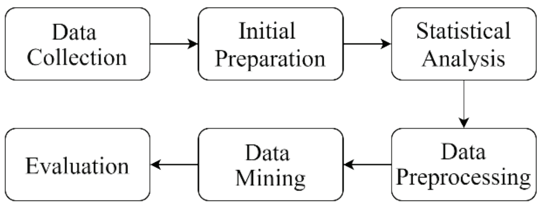

Fig.1. EDM Stages Framework

The data preprocessing stage consists of two essential components: Data transformation and feature selection. Feature selection focuses on identifying a subset of relevant variables while eliminating less significant or redundant ones. This process enhances the accuracy of predictive models and optimizes computational efficiency by reducing processing time. The DM process involves creating various types of models to evaluate and select the most appropriate model to summarize the research results. A confusion matrix is commonly used to determine the performance of these models and provides a detailed evaluation of their accuracy and effectiveness.

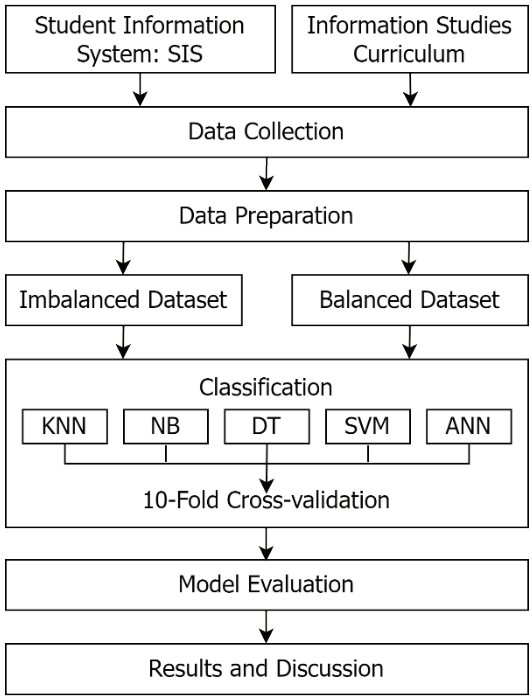

Fig.2. Research Framework

Fig.2 presents our research framework relevant to the EDM process. In the DM step, we create a model for predicting academic achievement using supervised ML to estimate the expected values of dependent variables based on the characteristics of the independent variables. Classification is the most popular method, and the most common classification algorithms are the Bayesian method, Neural Network, and Decision Tree. Apart from those three well-known algorithms, this study also includes the K-Nearest Neighbor and the Support Vector Machine for our model formation. As for data mining tools, WEKA is the most used tool for building predictive models since it has the functionality to answer all types of DM problems. RapidMiner is another widely adopted tool, ranking as the second most popular in the field [21].

B. Data Preparation

The volume of data is expanding rapidly, benefiting various academic disciplines, particularly in decision-making processes supported by computer technology and information management. The IS program is one of the key disciplines focused on information management through the application of modern technologies. It integrates the theories of information management and information technology. The program is designed to equip students with the knowledge and skills required to work in information-centric organizations, such as information agencies, libraries, and the broader information technology sector. This study utilizes data from the IS students of Burapha University, in conjunction with the IS curriculum.

Our study employs secondary data sourced from the SIS. The dataset is categorized into three types of variables: Demographic information, course grades, and early-stage GPA from the first six semesters. It includes data from 275 students who graduated from the Department of Information Studies between 2020 and 2023. Demographic information encompasses details such as students' home region, gender, number of siblings, parents' occupation and income, and high school GPA (SGPA). These variables are analyzed to identify patterns and factors influencing academic achievement among IS students.

This study investigates students' academic achievement over their four-year study period. The courses are categorized into three fundamental areas: General Education (GE), Information Science (IS), and Information Technology (IT). The IS and IT courses are further divided into core and elective courses, allowing a detailed analysis of students' academic performance. The dataset includes 12 GE courses, 18 core courses, and 10 elective courses, with each category split equally between IS and IT courses. The dataset is in its original form and prepared for the initial data preparation phase. Table I presents the variables used in the predictive models.

TABLE I
VARIABLES IN THE DATASETS

Type	Variables	Domain
Demographic	SGPA	{Excellent, Very Good, Good, Fair}
	Region	{North, South, Northeast, East, Central}
	Gender	{Male, Female}
	Sibling	{Yes, No}
	Father Income	{High, Medium, Low, None, Undefine}
	Father Occupation	{Government Officer, State Enterprise, Private Employee, Personal Business, Agriculture, Undefined}
	Mother Income	{High, Medium, Low, None, Undefined}
	Mother Occupation	{Government Officer, State Enterprise, Private Employee, Personal Business, Agriculture, Undefined}
Academic	GE Courses	{A, B+, B, C+, C, D+, D}
	IS Courses	{A, B+, B, C+, C, D+, D}
	IT Courses	{A, B+, B, C+, C, D+, D}
	GPA1-GPA6	{Excellent, Very Good, Good, Fair}
	AGPA2-AGPA6	{Excellent, Very Good, Good, Fair}

Academic achievement upon graduation is categorized into four classes based on the final GPA: Excellent, Very Good, Good, and Fair, as shown in Table II. After removing noise and missing values, the dataset includes information from 263 students out of an initial 275. This data cleaning improves the efficiency of the classification model and simplifies the computational process. Furthermore, to address class imbalance, the Synthetic Minority Over-Sampling Technique (SMOTE) was applied to generate additional samples, expanding the dataset from 263 to 400 instances.

TABLE II
TARGET CLASSES

GPA	Classes	No. of Instances (Original Dataset)	No. of Instances (Balanced Dataset)
3.50 – 4.00	Excellent	35	100
3.00 – 3.49	Very Good	102	100
2.50 – 2.99	Good	100	100
2.00 – 2.49	Fair	26	100
Total		**263**	**400**

An imbalanced dataset occurs when the instances of the target class are not evenly distributed. In this study, the dataset contains 35 samples for the "Excellent" class, 102 for "Very Good", 100 for "Good", and 26 for "Fair". The unequal distribution of instances in these classes results in an imbalanced dataset [22], [23], [24], which could potentially lead to inaccurate results. To address this issue, the SMOTE is applied to oversample the minority classes. As a result, the original dataset of 263 instances is expanded to 400 instances, with each target class containing 100 instances, thereby ensuring a balanced distribution across all classes.

Feature selection [25], [26] is a crucial pre-processing step in the DM process, aiming to identify and rank the importance of variables. Various ranking techniques are classified into filter-based, wrapper-based, and hybrid methods. In our study, we adopted a wrapper-based method, using the best classifier identified in our experiments to rank the variables based on their classification accuracy. The variable with the highest accuracy is considered the strongest correlation with the final academic performance.

C. Prediction Models

EDM models are broadly categorized into two types: predictive and descriptive models. Predictive models are employed to forecast outcomes using supervised learning techniques, while descriptive models aim to identify patterns that describe the underlying structure and relationships within unsupervised data. In this study, predictive models are applied to the datasets, utilizing five widely recognized ML algorithms: K-Nearest Neighbor (KNN), Naive Bayes (NB), Decision Tree (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN). These algorithms are among the most used in EDM research [27], [28]. The selection of these models for implementation is driven by the goal of identifying the most accurate predictive approach.

1) *K-Nearest Neighbor (KNN)* – KNN is a method of classifying data by comparing the data with sample data in the dataset. It refers to Instance-based Learning, which means the learned data is stored as an instance in the dataset. While classifying the new data, KNN will find K samples of data in the dataset that most resemble the new data (neighbors) and use them together to decide what class of new data should be. The principle behind this algorithm is to find the similar characteristics of the new data with the nearby dataset. Deciding the class of new data can be done using a method called "Majority Vote"; that is, new data will be the same class as the larger number of the neighbor data. Euclidean distance is a common measurement used to calculate the distance between data points. Classification using the KNN method can provide much accuracy depending on many factors, such as the completeness of the sample data used to represent the entire data or how much noise is in the data. Selecting a small value of K, such as $K = 1$ or 2, may cause misclassification because the closest data may be noise. Therefore, the value of K should be defined appropriately. This study has selected $K = 5$ for the experiments.

2) *Naive Bayes (NB)* – The NB method is a fundamental classification technique that utilizes probability theory based on Bayes' theorem. It classifies data into predefined groups by applying probabilistic principles. The NB algorithm assigns each data instance to the class with the highest posterior probability, ensuring optimal classification based on the given probabilistic framework.

3) *Decision Tree (DT)* – The DT method is a classification technique based on the concept of divide and conquer. Initially, the dataset is divided into smaller parts based on the values of the variables. The collection of decision nodes is connected by branches extending from the root node to the leaf node. Each node contains a condition that uses one of the data variables to decide on one child node. Decision-making starts at the root node and then moves on to the child nodes until reaching the leaf nodes, which are class nodes. The depth of the trees is related to how fast the model can make decisions. The selection of variables and conditions must be justified to obtain a tree that can classify the data as accurately as possible. This study uses the C4.5 algorithm to build the DT model.

4) *Support Vector Machine (SVM)* – SVM is a supervised learning algorithm designed for building classification models, particularly well-suited for datasets characterized by small sample sizes and a high dimensionality of features. The fundamental concept of SVM revolves around the creation of decision boundaries, referred to as hyperplanes, which partition the feature space to distinguish between different classes. The primary objective of SVM is to identify the optimal hyperplane by maximizing the margin, which is defined as the aggregate of the shortest distances from the hyperplane to the nearest data points of each class. This approach ensures enhanced generalization and robust classification performance. Regarding the experiment, the SVM model is configured with a regularization parameter C set to 1.0, balancing the trade-off between maximizing the margin and minimizing classification errors. A polynomial kernel is selected to capture complex and nonlinear relationships within the data. Normalization is applied to ensure that the features are on a consistent scale, which enhances the model's convergence and stability. The tolerance for the stopping criterion is 0.001, allowing the training process to terminate once improvements fall below this threshold.

5) *Artificial Neural Network (ANN)* – ANN represents a branch of AI whose structure and functionality resemble the neural networks of biological organisms. This technique is suitable for a non-linear fitting method. The ANN Model adjusts itself in response to the input associated with the learning rules. The feed-forward network, which consists of multiple layers of neurons, provides parameter calculation in several iterations to get the best configuration. The neurons' connections between layers are fully connected. Each neuron sends its computation results to every neuron in the next layer. Each link between neurons consists of a weight value that magnifies the value and passes it over its link by multiplying it by this weight. The result is transmitted to the next layer of neurons. Repeat this process from the input to the output layer.

In the experiment, the ANN model is initialized using default parameters included in the mining tool. The hidden layer is dynamically optimized by automatically adjusting the number of neurons based on the input data. A learning rate of 0.3 regulated the weight updates during training, while a momentum coefficient of 0.2 is incorporated to enhance convergence speed and minimize oscillations in the gradient descent process. The training process is conducted over 500 iterations to ensure adequate learning and model stability.

D. Models Evaluation

A common technique to evaluate the goodness of the DM algorithm is to apply the confusion matrix shown in Table III.

TABLE III
CONFUSION MATRIX

Observation	Positive	Negative
Predicted Positive	True Positive (TP)	False Positive (FP)
Predicted Negative	False Negative (FN)	True Negative (TN)

The class value of True Positive (TP) has the predicted class as YES and is YES, while the class value of False Negative (FN) has the predicted class as NO and is YES. Similarly, False Positive (FP) has the predicted class as YES and is NO, while True Negative (TN) has the predicted class as NO and is actually NO. These terms (TP, FN, FP, TN) represent frequency values and are used to construct a confusion matrix, which is a valuable tool for evaluating the performance of predictive models. The confusion matrix enables the calculation of various performance metrics, as shown in Table IV.

TABLE IV
MEASUREMENT

Performance Criteria	Formula
Accuracy	(TP+TN) / (TP+TN+FP+FN)
Precision	TP / (TP+FP)
Recall	TP / (TP+FN)
F-Measure	2 x (Precision x Recall) / (Precision + Recall)

Model evaluation is a critical step in assessing the effectiveness of a predictive model. To identify the best-performing model, it is necessary to compare multiple models using various evaluation metrics. Accuracy is the performance criterion that is relevant to all parameters. The higher the TP and TN rates, the more accurate the prediction is, which means better model performance. On the other hand, high FP and FN rates present a signal of incorrect prediction. Regarding precision and recall, the TP rate directly affects the performance. Accordingly, we focus on the TP rate and the classification accuracy to get the optimal prediction.

IV. RESULTS

In the experiment, the dataset is partitioned into two subsets: A training set and a testing set. Using the training set to build models by applying various algorithms, while the testing set is employed for model validation. Our EDM process incorporates the *K*-fold cross-validation technique, with the value of *K* set to 10. This choice of *K* is widely used in EDM research due to its effectiveness in balancing bias and variance. Weka is selected as the primary tool due to its comprehensive functionality for implementing and evaluating data mining models.

A. Demographic Information

Table V illustrates that academic achievement predictions based on demographic information have minimal influence on overall performance across all classifiers. The highest accuracy for the imbalanced dataset is achieved using the DT model, highlighted in bold, with an accuracy of 39.2%. Other models yield slightly lower accuracies, with the NB model performing the worst at 33.5% accuracy. When using a balanced dataset, the impact of demographic information on prediction accuracy remains limited, regardless of the model applied. The best-performing model in this case is the ANN model, achieving an accuracy of 58%, while other models produce slightly lower results. Similar to the imbalanced dataset scenario, the NB model demonstrates the lowest accuracy, at 51.8%. These findings suggest that demographic data contribute minimally to academic performance prediction, indicating that demographic attributes are not a significant factor for effective classification.

TABLE V
PERFORMANCE BASED ON
DEMOGRAPHIC INFORMATION

Classifiers	Accuracy (%)	
	Imbalanced	Balanced
KNN	36.9	52.3
NB	33.5	51.8
DT	**39.2**	56.3
SVM	35.4	57.5
ANN	35.7	**58.0**

In Table VI, the variable prioritization results for demographic datasets are analyzed using a wrapper-based technique. The DT classifier is applied to the imbalanced dataset, while the ANN is used for the balanced dataset, as both classifiers yield the highest accuracy. The analysis reveals that the SGPA achieves the highest accuracy, particularly on the balanced dataset, with a score of 45.8%. In contrast, the remaining demographic factors demonstrate significantly lower accuracy scores. Consequently, variables related to demographic data are found to have minimal impact on student performance and are deemed insignificant in predicting academic outcomes.

TABLE VI
VARIABLE IMPORTANCE BASED ON DEMOGRAPHIC INFORMATION USING DT FOR IMBALANCED DATASET AND ANN FOR BALANCED DATASET

Order	Variables	Accuracy (%) (Imbalanced)	Accuracy (%) (Balanced)
1	SGPA	41.1	45.8
2	Father Occupation	40.7	28.0
3	Mother Occupation	39.5	25.0
4	Mother Income	38.8	34.5
5	Region	38.4	26.0
6	Father Income	37.3	26.3
7	Gender	37.3	27.5
8	Sibling	36.1	32.3

B. Grade Information

This section evaluates the classification performance across three datasets: General Education (GE) courses (12 courses), core courses (18 courses), and elective courses (10 courses). Additionally, the final column presents the classification results based on all courses combined (40 courses), as shown in Table VII.

TABLE VII
PERFORMANCE BASED ON GRADES DATASETS FROM EACH COURSE

	Classifiers	Accuracy (%)			
		GE	Core	Elec.	All
Imbalanced	KNN	59.7	67.7	62.0	73.8
	NB	67.3	**81.4**	74.5	84.4
	DT	53.6	59.3	56.7	61.6
	SVM	63.1	68.8	65.8	76.8
	ANN	61.6	74.5	68.1	80.6
Balanced	KNN	68.0	76.3	74.3	78.0
	NB	75.8	**87.8**	84.8	91.3
	DT	69.0	74.5	72.5	75.0
	SVM	77.3	79.0	81.8	85.0
	ANN	76.5	84.3	81.3	87.8

The reported results represent the estimated accuracy of five models applied to both balanced and imbalanced datasets. For the imbalanced dataset, the NB model achieves the highest accuracy of 81.4% when applied to core courses. However, the balanced dataset yields improved accuracy, reaching 87.8% using the same NB model. In contrast, elective and GE courses exhibit lower predictive performance compared to core courses. Among all the classification techniques, the DT model demonstrates the lowest accuracy. The implementation of the SMOTE enhances classification accuracy, indicating its effectiveness in addressing data imbalance. This suggests that institutions can leverage this technique to classify students' academic performance and formulate policies for improvement. The NB classifier consistently delivers optimal results and outperforms other models when applied to the balanced dataset.

Further analysis is conducted by separating core and elective courses into Information Science (IS) and Information Technology (IT) datasets. The datasets are divided into IS core and IT core, each consisting of 9 courses. Similarly, the elective courses are split into IS elective and IT elective datasets, each containing 5 courses.

TABLE VIII
PERFORMANCE BASED ON GRADES FROM IS AND IT COURSE

	Classifiers	Accuracy (%)			
		IS Core	IT Core	IS Elec.	IT Elec.
Imbalanced	KNN	69.6	65.8	55.1	60.1
	NB	**75.3**	74.5	61.2	68.4
	DT	64.3	59.3	49.8	59.3
	SVM	70.3	65.0	64.3	65.0
	ANN	72.6	63.9	55.9	63.9
Balanced	KNN	77.5	73.5	66.5	72.3
	NB	**85.8**	83.0	76.0	80.0
	DT	77.5	74.0	68.8	74.0
	SVM	83.0	78.5	76.3	79.3
	ANN	80.5	79.0	71.5	76.3

The results presented in Table VIII align with the findings in Table VII, indicating that the NB model achieves the highest accuracy of 85.3% when applied to IS core courses using the balanced dataset. This suggests that IS courses have a greater influence on the target class compared to IT courses across all classifiers. However, when considering only elective courses, IT electives exhibit better predictive performance than IS electives. Since elective courses are typically taken in the later stages of study, greater emphasis is placed on core courses. Experimental findings suggest that core courses consistently yield the highest classification accuracy for both balanced and imbalanced datasets. Among all classifiers, the NB model demonstrates the best performance, making it the most effective classification technique for predicting academic achievement based on course performance.

TABLE IX
VARIABLE IMPORTANCE BASED ON GRADE FROM IS COURSES

Order	Courses	Accuracy (%)
1	Organization of Information Resources	63.5
2	Information and Reference Services	61.8
3	Library of Congress Classification	60.3
4	Library Automation Systems	59.8
5	Cataloging of Information Resources	59.3
6	Information Science	56.5
7	Reading for Information Professional	54.3
8	Collection Development	53.8
9	Management of Information Institutes	48.3

Additionally, Tables IX and X provide a ranking of core courses using the NB classifier as an indicator. The classification of IS and IT courses is based on course descriptions. IS courses primarily focus on information management, whereas IT courses emphasize the application of modern technology to information-related tasks, encompassing principles, theories, and software tools. Among the IS courses, *Organization of Information Resources* has the highest predictive impact, achieving an accuracy of 63.5%. This is followed by *Information and Reference Services* and *Library of Congress*, which yield accuracies of 61.8% and 60.3%, respectively. These findings highlight the significance of IS courses in predicting academic performance, emphasizing their role in shaping students' overall achievement.

TABLE X
VARIABLE IMPORTANCE BASED ON GRADE FROM IT COURSES

Order	Courses	Accuracy (%)
1	Research and Statistics in Information Studies	70.3
2	Information Systems Analysis and Design	63.8
3	Electronic Information and Record Management	61.5
4	Programming in Information Work	55.8
5	Information Technology	53.8
6	Database Management for Information Work	51.3
7	Web Design for Information Work	51.3
8	Seminar on Current Issues and Trend in Information Science	44.0
9	Presentation and Training in Information Work	43.3

On the other hand, the IT dataset includes 9 compulsory courses. Among them, *Research and Statistics in Information Studies* and *Information Systems Analysis and Design* demonstrate superior predictive performance compared to IS courses, achieving accuracies of 70.3% and 63.8%, respectively. Our analysis suggests that individual IT courses exert a stronger influence on academic achievement prediction. However, when considering the entirety of courses within each dataset, IS courses exhibit a closer correlation with the target outcomes. This indicates that while specific IT courses contribute significantly to performance prediction, the overall impact of IS courses remains more substantial in determining academic success.

C. Early-stage GPA Information

Table XI presents a comparison of the five models, incorporating factors such as GPA from the first semester to the sixth semester (GPA1-GPA6) and the average GPA (AGPA) from the second to the sixth semester (AGPA2-AGPA6). Given that each academic year comprises two semesters, the analysis considers GPA information up to the end of the third year. The results indicate that the NB classifier achieves the highest accuracy of 76.8% for GPA4 when applied to the imbalanced dataset. Additionally, GPAs and AGPAs yield consistent performance across all predictive models throughout the six semesters. For AGPA datasets, classification accuracy steadily improves from the end of the second semester to the end of the sixth semester across all classification techniques. Furthermore, in the balanced dataset, GPA5 achieves the highest accuracy of 77.8% using the NB classifier. Therefore, the findings suggest that the NB model provides the most effective prediction, with GPA4 being the best predictor for the imbalanced dataset and GPA5 for the balanced dataset.

TABLE XI
PERFORMANCE BASED ON EARLY-STAGE GPA FROM EACH SEMESTER

	GPA	Accuracy (%)				
		KNN	NB	DT	SVM	ANN
Imbalanced	GPA1	59.7	59.7	59.7	59.7	57.4
	GPA2	64.3	64.3	64.3	64.0	64.3
	GPA3	62.4	62.4	62.4	62.4	62.4
	GPA4	71.1	**76.8**	71.1	71.1	71.1
	GPA5	68.4	68.4	68.4	68.4	68.4
	GPA6	64.3	64.0	64.3	64.3	65.8
	AGPA2	60.1	60.1	60.1	60.1	59.3
	AGPA3	62.4	63.5	63.5	63.5	63.5
	AGPA4	65.0	64.3	65.0	65.0	65.8
	AGPA5	66.5	65.0	66.5	66.5	66.9
	AGPA6	72.2	72.2	72.2	72.2	72.2
Balanced	GPA1	57.5	57.5	57.5	56.0	59.5
	GPA2	58.0	58.0	58.0	58.0	58.0
	GPA3	63.5	63.5	63.5	63.5	62.5
	GPA4	74.0	74.0	74.0	74.0	74.0
	GPA5	**77.8**	**77.8**	**77.8**	**77.8**	**77.8**
	GPA6	76.0	76.0	76.0	76.0	76.0
	AGPA2	62.0	62.0	62.0	62.0	62.0
	AGPA3	62.8	63.5	63.5	63.5	63.5
	AGPA4	68.5	68.5	68.5	68.5	68.5
	AGPA5	68.8	68.8	68.8	68.8	68.8
	AGPA6	71.0	71.0	71.0	71.0	71.0

TABLE XII
VARIABLE IMPORTANCE BASED ON EARLY-STAGE GPA USING NB CLASSIFIER

Order	Imbalanced		Balanced	
	Variables	Accuracy (%)	Variables	Accuracy (%)
1	GPA4	76.8	GPA5	77.8
2	AGPA6	72.2	GPA6	76.0
3	GPA5	68.4	GPA4	74.0
4	AGPA5	65.0	AGPA6	71.0
5	GPA2	64.3	AGPA5	68.8
6	AGPA4	64.3	AGPA4	68.5
7	GPA6	64.0	GPA3	63.5
8	AGPA3	63.5	AGPA3	63.5
9	GPA3	62.4	AGPA2	62.0
10	AGPA2	60.1	GPA2	58.0
11	GPA1	59.7	GPA1	57.5

The ranking of variables using GPA information, as determined by the NB classifier, is presented in Table XII. The GPA from the fifth semester, based on the balanced dataset, yields the highest accuracy compared to other variables. Factors related to GPA and AGPA consistently align with previous findings, demonstrating that as GPA or AGPA values approach the completion of academic studies, the prediction accuracy increases. This indicates that predictions become more reliable closer to graduation. However, this study's objective is to provide institutions with early insights into student performance before course completion. The results suggest that academic achievement can be estimated with an accuracy of 77.8% using GPA from the fifth semester (GPA5), offering a valuable tool for early intervention and support.

TABLE XIII
PERFORMANCE BASED ON THE INCREMENTAL INCLUSION OF GPA FROM EACH SEMESTER

	GPA	Accuracy (%)				
		KNN	NB	DT	SVM	ANN
Imbalanced	GPA1	**59.7**	**59.7**	**59.7**	**59.7**	57.4
	GPA2	63.1	63.5	63.9	**65.8**	64.6
	GPA3	68.4	72.2	**73.8**	71.5	70.3
	GPA4	77.9	**79.1**	76.4	75.7	76.0
	GPA5	81.4	**83.7**	73.8	81.4	79.5
	GPA6	83.3	**85.6**	75.7	84.4	82.1
Balanced	GPA1	57.5	57.5	57.5	56.0	**59.5**
	GPA2	63.8	**65.5**	64.5	64.3	65.0
	GPA3	74.0	75.3	76.5	**77.5**	76.3
	GPA4	81.0	83.5	**84.3**	80.0	83.5
	GPA5	83.5	**86.5**	82.3	84.5	81.5
	GPA6	87.3	**90.5**	86.0	89.8	87.3

Table XIII presents the results obtained by incrementally adding GPA datasets from GPA1 to GPA6. For instance, GPA1 consists of a single factor, while GPA2 includes both GPA1 and GPA2. Similarly, GPA3 comprises GPA1, GPA2, and GPA3, continuing this pattern until GPA6 incorporates all preceding GPA values. The results indicate that adding consecutive GPAs leads to an improvement in accuracy. Especially, the performance of the model on the balanced dataset shows a steady increase in accuracy from GPA1 to GPA6, reaching a maximum accuracy of 90.5%. While these outputs differ slightly from those of the AGPA datasets in Table XI, both results follow a similar trend, where predictive accuracy improves as the dataset includes more GPA information. The NB classifier achieved the highest accuracy, indicating that its combination with SMOTE enhances prediction performance.

V. DISCUSSION

The results reveal that student demographics have a negligible impact on academic performance. The most relevant factor is SGPA, which can be a useful indicator. Among the least influential factors are gender and siblings. Meanwhile, things like parents' occupation and incomes, or the religion of the students, have an even smaller effect. Hence, a student's background has little influence on their final academic achievement.

Grade information for GE courses demonstrated strong predictive accuracy, with the SVM model achieving 77.3% accuracy, followed by the ANN model at 76.5%. Therefore, the GE course grade data is valuable for prediction.

When analyzing courses in the fields of IS and IT, we discovered that among the core courses, the IS group demonstrated slightly higher predictive accuracy than the IT group, with accuracy rates of 85.8% and 83%, respectively, using the same NB model. Conversely, for elective courses, the IT group performed slightly better than the IS group, achieving 80% accuracy with the NB model, compared to 76.3% accuracy with the SVM model. Overall, those two groups exhibited similar performance, which is beneficial to the curriculum design by considering the courses between both IS and IT.

Early-stage GPA information provides highly accurate predictions and is a valuable dataset for forecasting final academic achievement. In the first three years, the NB model achieved a prediction accuracy of 90.5%, followed by the SVM model with 89.8% accuracy. This type of dataset is especially useful because it covers the earlier stage of the educational journey, unlike individual course grades, which may vary depending on the semester in which students enroll. Some courses are taken later in the program making them less useful for early predictions. When examining GPA per semester, the GPA5 shows the most promising results, with 77.8% accuracy. Analyzing cumulative GPA over the first six semesters further increases the accuracy score. Interestingly, however, the accuracy of cumulative GPA predictions is lower than that of semester-based GPA predictions. This is likely because the GPA of the first year are not a strong predictor for long-term academic success since students were in the process of adapting to get familiar with university life.

The reason that NB outperforms the other models is due to its special characteristics. NB naturally handles categorical features without complex preprocessing, which may complicate other models. In addition, NB works well with the independence feature that has small datasets. Its probabilistic nature and low complexity make it less prone to overfitting than models like decision tree or neural network. For instance, to predict whether a student will be categorized into which class, NB may achieve higher accuracy than more complex models when the dataset is small, imbalanced, or contains mostly categorical data like the dataset used in this study.

The predictive insights derived from student performance data can serve as a foundation for evidence-based educational interventions. For example, if we can identify low-achievement students early through the model, institutions can focus on supporting mechanisms, such as tutoring or academic counseling, to help that group of students. Furthermore, consistent underperformance in specific courses can indicate a weakness in the curriculum. Curriculum committees may have to revise the prerequisite course structure or integrate supplemental instruction into difficult courses. At the strategic level, this knowledge empowers institutions in making decisions for allocating resources more effectively, prioritizing academic support services, and implementing retention strategies. This not only enhances student success rates but also contributes to long-term outcomes such as graduation rates and institutional accountability.

The five different ML models yielded varying results. Some models performed well with certain datasets but not with others, which is why we need to apply multiple models to determine the most effective prediction. Among the models, NB consistently delivered the highest accuracy across most types of datasets. Meanwhile, ANN and SVM performed well but slightly lower than NB. DT and KNN models showed minor accuracy compared to the others. Therefore, regarding this particular case study, if a single predictive model were chosen, NB would be the best option for predicting the final academic performance on graduation.

VI. Conclusion

This study has explored factors that affect the students' academic achievement using EDM techniques. We applied various datasets, including demographic information, course grades, and early-stage GPA from each semester, in the experiments. Course grades were categorized into three areas: GE, IS, and IT. Moreover, we discovered the distinction between core and elective courses within the area of information science and information technology.

To address the imbalanced dataset, the SMOTE technique was applied to balance the number of instances across each class. We generated classification models using five different classifiers: KNN, NB, DT, SVM, and ANN. The results demonstrate that early-stage GPA provides the highest predictive accuracy among other factors. GPA from the fifth semester had the most significant impact on the prediction. Among the classifiers, the NB model outperforms the others. Among the various types of datasets analyzed, demographic information demonstrated a relatively moderate impact on the performance. The highest accuracy achieved using this data type was 58.0%, obtained with the ANN algorithm, indicating its limited effectiveness compared to other dataset types. On the other hand, grades from core courses with the NB algorithm produced a significant effect on the learning outcomes by achieving an accuracy of 87.8%. The IS core courses provide better predictions than the IT core courses, while IT elective courses slightly underperform compared to their IS counterparts.

The incremental inclusion of early-stage GPA data from GPA1 to GPA6 on the balanced dataset yielded the highest accuracy, reaching 90.5%. This indicates that having access to GPA information from additional early semesters enhances the accuracy of predicting academic achievement. The application of the SMOTE technique to balance the datasets led to improved performance across all dataset types. Consequently, EDM offers meaningful insights that facilitate early intervention, guide curriculum enhancement, and support student counseling efforts to raise final academic outcomes.

Future research is recommended to focus on integrating additional data sources, such as questionnaires, online learning activities, and other educational engagement metrics, to enhance predictive accuracy. Furthermore, exploring a wider range of machine learning algorithms, including Random Forest, Deep Neural Networks, Linear Regression, and Evolutionary Algorithms, could provide deeper insights and enhance the prediction models. These suggestions are capable of improving student performance prediction and supporting more effective academic decision-making. This research approach, which employs EDM techniques, can be applied to educational datasets with similar characteristics across various curriculum types.

Acknowledgment

This research was funded and supported by the Faculty of Humanities and Social Sciences, Burapha University.

References

[1] R. S. J. D. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *J. Educ. Data Min*., vol. 1, no. 1, pp. 3-16, Jan. 2009, https://doi.org/10.5281/zenodo.3554657

[2] V. L. Migueis, A. Freitas, P. J. V. Garcia, and A. Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," *Decis. Support Syst.*, vol. 115, pp. 36-51, Nov. 2018, https://doi.org/10.1016/j.dss.2018.09.001

[3] M. Anoopkumar and A. M. J. M. Z. Rahman, "A Review on Data Mining techniques and factors used in Educational Data Mining to predict student amelioration," in *Proc. 2016 Int. Conf. Data Min. Adv. Comput. (SAPIENCE),* 2016, pp. 122-133, https://doi.org/10.1109/SAPIENCE.2016.7684113.

[4] W. Xing, "Exploring the influences of MOOC design features on student performance and persistence," *Distance Educ*., vol. 40, no. 1, pp. 98-113, Dec. 2019, https://doi.org/10.1080/01587919.2018.1553560

[5] J. D. Parker, M. J. Hogan, J. M. Eastabrook, A. Oke, and L. M. Wood, "Emotional intelligence and student retention: Predicting the successful transition from high school to university," *Pers. Individ. Diffe*r., vol. 41, no. 7, pp. 1329-1336, Nov. 2006, https://doi.org/10.1016/j.paid.2006.04.022

[6] A. Richard-Eaglin, "Predicting student success in nurse practitioner programs," *J. Am. Assoc. Nurse Pract.*, vol. 29, no. 10, pp. 600-605, Oct. 2017, https://doi.org/10.1002/2327-6924.12502

[7] E. Alqurashi, "Predicting student satisfaction and perceived learning within online learning environments," *Distance Educ.*, vol. 40, no. 1, pp. 133-148, Dec. 2018, https://doi.org/10.1080/01587919.2018.1553562

[8] B. Pérez, C. Castellanos, and D. Correal, "Predicting student drop-out rates using data mining techniques: A case study," in *Proc. Int. Conf. Comput. Sci. Appl*., 2018, pp. 111-125, https://doi.org/10.1007/978-3-030-03023-0_10

[9] A. M. Shahiri, W. Husain, and N. A. Rashid, "A review on predicting student's performance using data mining techniques," *Procedia Comput. Sci.*, vol. 72, pp. 414-422, 2015, https://doi.org/10.1016/j.procs.2015.12.157

[10] H. Almarabeh, "Analysis of students' performance by using different data mining classifiers," *Int. J. Mod. Educ. Comput. Sci.*, vol. 9, no. 8, pp. 9-15, Aug. 2017, https://doi.org/10.5815/ijmecs.2017.08.02

[11] A. Raheela, M. Agathe, A. A. Syed, and G. H. Najmi, "Analyzing undergraduate students' performance using educational data mining," *Comput. Educ.*, vol. 113, pp. 177-194, Oct. 2017, https://doi.org/10.1016/j.compedu.2017.05.007

[12] Y. Baashar et al., "Toward Predicting Student's Academic Performance Using Artificial Neural Networks (ANNs)," *Appl. Sci.*, vol. 12, no. 3, pp. 1-16, Jan. 2022, https://doi.org/10.3390/app12031289

[13] Y. Baashar et al., "Evaluation of postgraduate academic performance using artificial intelligence models," *Alexandria Eng. J*., vol. 61, no. 12, pp. 9867-9878, Dec. 2022, https://doi.org/10.1016/j.aej.2022.03.021

[14] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis," *Telemat. Inform*., vol. 37, pp. 13-49, Apr. 2019, https://doi.org/10.1016/j.tele.2019.01.007

[15] B. Owaidat, "Exploring the Accuracy and Reliability of Machine Learning Approaches for Student Performance," *Appl. Comput. Sci.*, vol. 20, no. 3, pp. 67-84, Sep. 2024, https://doi.org/10.35784/acs-2024-29

[16] S. Sarker, M. K. Paul, S. T. H. Thasin, and M. A. M. Hasan, "Analyzing students' academic performance using educational data mining," *Comput. Educ.: Artif. Intell.*, vol. 7, p. 100263, Dec. 2024, https://doi.org/10.1016/j.caeai.2024.100263

[17] J. Zimmermann, K. H. Brodersen, H. R. Heinimann, and J. M. Buhmann, "A model-based approach to predicting graduate-level performance using indicators of undergraduate-level performance," *J. Educ. Data Min*., vol. 7, no. 3, pp. 151-176, Oct. 2015, https://doi.org/10.5281/zenodo.3554733

[18] S. Alex, J. J. V. Nayahi, and S. Kaddoura, "Deep convolutional neural networks with genetic algorithm-based synthetic minority over-sampling technique for improved imbalanced data classification," *Appl. Soft Comput.,*vol. 156, 111491 May 2024, https://doi.org/10.1016/j.asoc.2024.111491

[19] S. A. Alex, "Classification of imbalanced data using SMOTE and autoencoder based deep convolutional neural network," *Int. J. Uncertain. Fuzziness Knowl. Based Syst.,* vol. 31, no. 3, pp. 437-469, 2023. https://doi.org/10.1142/s0218488523500228

[20] E. Alyahyan and D. Dustegor, "Predicting academic success in higher education: Literature review and best practices," *Int. J. Educ. Technol. High. Educ*., vol. 17, no. 3, pp. 1-21, Feb. 2020, https://doi.org/10.1186/s41239-020-0177-7

[21] S. Jayaprakash, "A Survey on academic progression of students in tertiary education using classification algorithms," *Int. J. Eng. Technol*., vol. 8, no. 6, pp. 111-115, Feb. 2019.

[22] W. Intayoad, C. Kamyod, and P. Temdee, "Synthetic minority over-sampling for improving imbalanced data in educational web usage mining," *ECTI Trans. Comput. Inf. Technol.*, vol. 12, no. 2, pp. 118-129, Feb. 2019, https://doi.org/10.37936/ecti-cit.2018122.133280

[23] A. AL-Ashoor and S. Abdullah, "Examining techniques to solving imbalanced datasets in educational data mining systems," *Int. J. Comput*., vol. 21, no. 2, pp. 205-213, Jun. 2022, https://doi.org/10.47839/ijc.21.2.2589

[24] S. Aliga, A. S. Gaafar, and A. K. Hamoud, "Predicting students performance using supervised machine learning based on imbalanced dataset and wrapper feature selection," *Informatica*, vol. 47, no. 1, pp. 11-20, 2023, https://doi.org/10.31449/inf.v47i1.4519

[25] K. Sutha and J. J. Tamilselvi, "A review of feature selection algorithms for data mining techniques," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 6, pp. 63-67, Jun. 2015.

[26] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70-79, Jul. 2018, https://doi.org/10.1016/j.neucom.2017.11.077

[27] M. Zafari, A. S. Niaraki, S. M. Choi, and A. Esmaeily, "A practical model for the evaluation of high school student performance based on machine learning," *Applied Sciences*, vol. 11, no. 23, p. 11534, Dec. 2021 https://doi.org/10.3390/app112311534

[28] M. R. Islam, A. M. Nitu, M. A Marjan, M. P. Uddin, M. I. Afjal, and M. A. A. Mamun, "Enhancing tertiary students' programming skills with an explainable educational data mining approach," *PLoS ONE* , vol. 19, no. 9, e0307536, Sep. 2024, https://doi.org/10.1371/journal.pone.0307536

Knitchepon Chotchantarakun is a lecturer in the Department of Information Studies, Faculty of Humanities and Social Sciences at Burapha University (BUU), Thailand. He received his Ph.D. in Computer Science from the Graduate School of Applied Statistics, National Institute of Development Administration (NIDA), Thailand, in 2021. He earned his M.Sc. in Computer Science from Chulalongkorn University (CU), Thailand, in 2006, and his B.Sc. in Computer Science with second-class honors from Mahidol University International College (MUIC), Thailand, in 2003. His research interests include evolutionary algorithms, feature selection, optimization, data mining, and machine learning.

PAPER FORMAT
(IEEE Style)

I. Format

- Your paper must use a paper size corresponding to A 4 which is 210 mm (8.27 inch) Wide and 297 mm (11.69 inch)
- Your paper must be in two column format
- Articles not more than 15 pages in length, single-sided A4 paper, margins (top, bottom, left, right) are 1 inch (2.54 cm)
- Abstract and References and content set to double columns,
- English font is Times New Roman, as follows:

TABLE I
Font Sizes for Papers

Content	Font Size	Labelling
Title (Single column)	18 (CT)	bold
Authors (Single column)	11 (CT)	bold
Authors Information (Single column)	10 (CT)	regular
Abstract	10 (LRJ)	bold
Index Terms (Keywords)	10 (LRJ)	bold
Content	10 (LRJ))	regular
Heading1	10 (CT)	regular (Capitalization)
Heading 2	10 (LJ)	regular
Table Title (Place above the Table)	8 (CT)	regular
Table content	8 (CT)	regular
Figure caption (Place below the figure)	8 (LJ)	regular
Reference Head	10 (CT)	regular (Capitalization)
Reference	8 (LJ)	regular
Author Profiles	10 (LRJ)	bold author name/ profile regular

CT=Centre Text, LJ=Left Justified, RJ=Right Justified, LRJ=Left & Right Justified

II. Composition of the Article

A. Article title

B. Authors information, Write (all) the author's name, affiliation, department, city, country and E-mail (set to Single Column) all.

C. Abstract, Must be under 200 words and not include subheadings or citations. Define all symbols used in the abstract. Do not delete the blank line immediately above the abstract.

D. Index Terms, Enter key words or phrases in alphabetical order, separated by commas.

E. Content

1) Academic article, should include: Introduction, Content, and Conclusion.

2) Research article, should include: introduction, literature review, Materials methods, Results, Discussion, and conclusion.

Clearly summarize the important findings of the paper. It should contain such as objectives, methods and major results.

F. Introduction

The Introduction section of reference text expands on the background of the work (some overlap with the Abstract is acceptable). The introduction should not include subheadings.

G. Pictures, table, etc., Must be use in numerical order in the article, provided the source correctly, cannot use other people' copyright.

Chart should be colored contrastingly or in black and white.

H. Reference

1) Cited in the main text. Indicate the number in the [ ] mark at the end of the text or the name of the referring person. Let the numbers be in the same line of content as [1].

2) Cited after the article. Put all bibliographical reference after articles, and order according to the author's name, please refer IEEE format. The footer reference format is as follows.

III. Rererences

References in research articles and scholarly articles. For academic and research journals, INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET). The technology defines referrals according to the IEEE format. All references should be listed at the end of the paper using the following.

Basic format for books:

J. K. Author, "Title of chapter in the book," in *Title of His Published Book,* xth ed. City of Publisher, Country if not USA: Abbrev. of Publisher, year, ch. x, sec. x, pp. xxx–xxx.

Examples:

[1] G. O. Young, "Synthetic structure of industrial plastics," in *Plastics,* 2nd ed., vol. 3, J. Peters, Ed. New York: McGraw-Hill, 1964, pp. 15-64.

[2] W.-K. Chen, *Linear Networks and Systems.* Belmont, CA: Wadsworth, 1993, pp. 123-135.

Basic format for periodicals:

J. K. Author, "Name of paper," *Abbrev. Title of Periodical*, vol. x, no. x, pp. xxx-xxx, Abbrev. Month. year.

Examples:

[3] J. U. Duncombe, "Infrared navigation—Part I: An assessment of feasibility," *IEEE Trans. Electron Devices*, vol. ED-11, no. 1, pp. 34-39, Jan. 1959.

[4] E. P. Wigner, "Theory of traveling-wave optical laser," *Phys. Rev.*, vol. 134, pp. A635-A646, Dec. 1965.

[5] E. H. Miller, "A note on reflector arrays," *IEEE Trans. Antennas Propagat.*, to be published.

Basic format for reports:

J. K. Author, "Title of report," Abbrev. Name of Co., City of Co., Abbrev. State, Rep. xxx, year.

Examples:

[6] E. E. Reber, R. L. Michell, and C. J. Carter, "Oxygen absorption in the earth's atmosphere," Aerospace Corp., Los Angeles, CA, Tech. Rep. TR-0200 (4230-46)-3, Nov. 1988.

[7] J. H. Davis and J. R. Cogdell, "Calibration program for the 16-foot antenna," Elect. Eng. Res. Lab., Univ. Texas, Austin, Tech. Memo. NGL-006-69-3, Nov. 15, 1987.

Basic format for handbooks:

Name of Manual/Handbook, x ed., Abbrev. Name of Co., City of Co., Abbrev. State, year, pp. xxx-xxx.

Examples:

[8] *Transmission Systems for Communications*, 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44-60.

[9] *Motorola Semiconductor Data Manual*, Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.

Basic format for books (when available online):

Author. (year, month day). *Title.* (edition) [Type of medium]. *volume (issue).* Available: site/path/file

Example:

[10] J. Jones. (1991, May 10). *Networks.* (2nd ed.) [Online]. Available: http://www.atm.com

Basic format for journals (when available online):

Author. (year, month). Title. *Journal.* [Type of medium]. *volume (issue),* pages. Available: site/path/file

Example:

[11] R. J. Vidmar. (1992, Aug.). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. *21(3),* pp. 876-880. Available: http://www.halcyon.com/pub/journals/21ps03-vidmar

Basic format for papers presented at conferences (when available online):

Author. (year, month). Title. Presented at Conference title. [Type of Medium]. Available: site/path/file

Example:

[12] PROCESS Corp., MA. Intranets: Internet technologies deployed behind the firewall for corporate productivity. Presented at INET96 Annual Meeting. [Online]. Available: http://home.process.com/Intranets/wp2.htp

Basic format for reports and handbooks (when available online):

Author. (year, month). Title. Comp any . City, State or Country. [Type of Medium]. Available: site/path/file

Example:

[13] S. L. Talleen. (1996, Apr.). The Intranet Archi-tecture: M anaging information in the new paradigm. Amdahl Corp., CA. [Online]. Available:http://www.amdahl.com/doc/products/bsg/intra/infra/html

Basic format for computer programs and electronic documents (when available online):

ISO recommends that capitalization follow the accepted practice for the language or script in which the information is given.

Example:

[14] A. Harriman. (1993, June). Compendium of genealogical software. *Humanist.* [Online]. Available e-mail: HUMANIST @NYVM.ORG Message: get GENEALOGY REPORT

Basic format for patents (when available online):

Name of the invention, by inventor's name. (year, month day). *Patent Number* [Type of medium]. Available: site/path/file

Example:

[15] Musical toothbrush with adjustable neck and mirror, by L.M.R. Brooks. (1992, May 19). *Patent D 326 189* [Online]. Available: NEXIS Library: LEXPAT File: DESIGN

Basic format for conference proceedings (published):

J. K. Author, "Title of paper," in *Abbreviated Name of Conf.*, City of Conf., Abbrev. State (if given), year, pp. xxxxxx.

Example:

[16] D. B. Payne and J. R. Stern, "Wavelength-switched pas- sively coupled single-mode optical network," in *Proc. IOOC-ECOC,* 1985, pp. 585-590.

Example for papers presented at conferences (unpublished):

[17] D. Ebehard and E. Voges, "Digital single sideband detection for interferometric sensors," presented at the 2nd Int. Conf. Optical Fiber Sensors, Stuttgart, Germany, Jan. 2-5, 1984.

Basic format for patents:

J. K. Author, "Title of patent," U.S. Patent x xxx xxx, Abbrev. Month. day, year.

Example:

[18] G. Brandli and M. Dick, "Alternating current fed power supply," U.S. Patent 4 084 217, Nov. 4, 1978.

Basic format for theses (M.S.) and dissertations (Ph.D.):

J. K. Author, "Title of thesis," M.S. thesis, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

J. K. Author, "Title of dissertation," Ph.D. dissertation, Abbrev. Dept., Abbrev. Univ., City of Univ., Abbrev. State, year.

Examples:

[19] J. O. Williams, "Narrow-band analyzer," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ., Cambridge, MA, 1993.

[20] N. Kawasaki, "Parametric study of thermal and chemical nonequilibrium nozzle flow," M.S. thesis, Dept. Electron. Eng., Osaka Univ., Osaka, Japan, 1993.

Basic format for the most common types of unpublished references:

J. K. Author, private communication, Abbrev. Month, year.

J. K. Author, "Title of paper," unpublished.

J. K. Author, "Title of paper," to be published.

Examples:

[21] A. Harrison, private communication, May 1995.

[22] B. Smith, "An approach to graphs of linear forms," unpublished.

[23] A. Brahms, "Representation error for real numbers in binary computer arithmetic," IEEE Computer Group Repository, Paper R-67-85.

Basic format for standards:

Title of Standard, Standard number, date.

Examples:

[24] IEEE Criteria for Class IE Electric Systems, IEEE Standard 308, 1969.

[25] Letter Symbols for Quantities, ANSI Standard Y10.5-1968.

First A. Author and the other authors may include biographies at the end of regular papers. Biographies are often not included in conference related papers. The first paragraph may contain a place and/or date of birth (list place, then date). Next, the author's educational background is listed. The degrees should be listed with type of degree in what field, which institution, city, state, and country, and year the degree was earned. The author's major field of study should be lower-cased.

The second paragraph uses the pronoun of the person (he or she) and not the author's last name. It lists military and work experience, including summer and fellowship jobs. Job titles are capitalized. The current job must have a location; previous positions may be listed without one. Information concerning previous publications may be included. Try not to list more than three books or published articles. The format for listing publishers of a book within the biography is: title of book (city, state: publisher name, year) similar to a reference. Current and previous research interests end the paragraph.

The third paragraph begins with the author's title and last name (e.g., Dr. Smith, Prof. Jones, Mr. Kajor, Ms. Hunter). List any memberships in professional societies. Finally, list any awards and work for committees and publications. If a photograph is provided, the biography will be indented around it. The photograph is placed at the top left of the biography, and should be of good quality, professional-looking, and black and white (see above example). Personal hobbies will be deleted from the biography. Following are two examples of an author's biography.

Second B. Author was born in Greenwich Village, New York City, in 1977. He received the B.S. and M.S. degrees in aerospace engineering from the University of Virginia, Charlottesville, in 2001 and the Ph.D. degree in mechanical engineering from Drexel University, Philadelphia, PA, in 2008. From 2001 to 2004, he was a Research Assistant with the Princeton Plasma Physics Laboratory. Since 2009, he has been an Assistant Professor with the Mechanical Engineering Department, Texas A&M University, College Station. He is the author of three books, more than 150 articles, and more than 70 inventions. His research interests include high-pressure and high-density nonthermal plasma discharge processes and applications, microscale plasma discharges, discharges in liquids, spectroscopic diagnostics, plasma propulsion, and innovation plasma applications. He is an Associate Editor of the journal *Earth*, *Moon*, *Planets*, and holds two patents.

Mr. Author was a recipient of the International Association of Geomagnetism and Aeronomy Young Scientist Award for Excellence in 2008, the IEEE Electromagnetic Compatibility Society Best Symposium Paper Award in 2011, and the American Geophysical Union Outstanding Student Paper Award in Fall 2005.

Third C. Author received the B.S. degree in mechanical engineering from National Chung Cheng University, Chiayi, Taiwan, in 2004 and the M.S. degree in mechanical engineering from National Tsing Hua University, Hsinchu, Taiwan, in 2006. He is currently pursuing the Ph.D. degree in mechanical engineering at Texas A&M University, College Station.

From 2008 to 2009, he was a Research Assistant with the Institute of Physics, Academia Sinica, Tapei, Taiwan. His research interest includes the development of surface processing and biological/ medical treatment techniques using nonthermal atmospheric pressure plasmas, fundamental study of plasma sources, and fabrication of micro- or nanostructured surfaces.

Mr. Author's awards and honors include the Frew Fellowship (Australian Academy of Science), the I. I. Rabi Prize (APS), the European Frequency and Time Forum Award, the Carl Zeiss Research Award, the William F. Meggers Award and the Adolph Lomb Medal (OSA).

Remark: More detail information, Pleases read Preparation of Papers for INTERNATIONAL SCIENTIFIC JOURNAL OF ENGINEERING AND TECHNOLOGY (ISJET), https://ph02.tci-thaijo.org/index.php/isjet/index

Panyapiwat Institute of Management (PIM)
85/1 Moo 2, Chaengwattana Rd,
Bang Talat, Pakkred, Nonthaburi 11120, Thailand
Tel. +66 2855 1560
https://www.tci-thaijo.org/index.php/isjet/index
https://isjet.pim.ac.th
E-mail: isjet@pim.ac.th