A publication of ISCA*:
International Society for Computers
and Their Applications

INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS

TABLE OF CONTENTS

Page

***"International Journal of Computers and Their Applications is Peer Reviewed".**

Volume 30, No. 3, Sept. 2023 **ISSN 1076-5204**

International Journal of Computers and Their Applications

A publication of the International Society for Computers and Their Applications

ISCA Headquaters • 278 Mankato Ave, #220, Winona, Minnesota 55987, USA
Phone 507-458-4517 • Email isca@isca-hq.org • URL: http://www.isca-hq.org

Editorial

It is my distinct honor, pleasure, and privilege to serve as the new Editor-in-Chief of the International Journal of Computers and Their Applications (IJCA) for the second year. I have a special passion for the International Society for Computers and their Applications. I have been a member of our society since 2014 and have served in various capacities. These have ranged from being on program committees of our conferences to being Program Chair of CATA 2021 and CATA 2022 and CATA 2023 and currently serving as one of the Ex-Officio Board Members. I am very grateful to the ISCA Board of Directors for giving me this opportunity to serve society and the journal in this role.

I would also like to thank all the editorial board, editorial staff, and the authors for their valuable contributions to the journal. Without everyone's help, the success of the journal would be impossible. I look forward to working with everyone in the coming years to maintain and further improve the journal's quality. I want to invite you to submit your quality work to the journal for consideration of publication. I also welcome proposals for special issues of the journal. If you have any suggestions to improve the journal, please feel free to contact me.

Dr. Ajay Bandi
School of Computer Science and Information Systems
Northwest Missouri State University
Maryville, MO 64468
Email: AJAY@nwmissouri.edu

In 2023, we have four issues planned (March, June, September, and December). The September issue includes the selected papers from CATA 2023 and open submissions. Ajay Bandi and Mohammad Hossain are the program co-chairs of CATA 2023. The September issue will contain the best papers from CATA 2023 and open submissions. The last issue is taking shape with a collection of submitted papers.

I would also like to announce that I will begin searching for a few reviewers to add to our team. There are a few areas in which we would like to strengthen our board. If you would like to be considered, please contact me via email with a cover letter and a copy of your CV.

Ajay Bandi, Editor-in-Chief
Email: AJAY@nwmissouri.edu

Guest Editorial
September 2023

This issue of the International Journal of Computers and their Applications (IJCA) is split into two parts. The first part is a collection of four refereed papers selected from CATA 2023, and the second part is IJCA-contributed papers that have gone through the normal review process. The papers in this issue cover a broad range of research interests in the community of computers and their applications.

ISCA Spring 2023 CATA Conference: CATA 2023 - The 38th International Conference on Computers and Their Applications, was held March 20-22, 2023. Due to the pandemic, it was held virtually. Each paper submitted to the CATA 2023 conference was reviewed by at least two international program committee members and by additional reviewers, judging for originality, technical contribution, significance, and quality of presentation. The proceedings for this conference can be found online at https://easychair.org/publications/volume/CATA2023

After the conference, the program committee members recommended the four best papers to be considered for publication in this special issue of IJCA. The authors were invited to submit a revised version of their papers. After extensive revisions and a second round of review, these papers were accepted for publication in this issue of the journal. The topics and main contributions of the papers are briefly summarized below:

LIAN KANG and PIERRE PAYEUR of the University of Ottawa, Canada presents their work "Object Recognition for Autonomous Vehicles from Combined Color and LiDAR Data". The paper discusses the importance of object detection in autonomous driving vehicles and the benefits of combining data from color cameras and LiDAR scanners. The proposed 3D object detector utilizing a bird's-eye view map and focal loss achieves a high speed of 46 frames per second with over 90% average precision. Furthermore, a more compact detector is introduced, maintaining a fast-processing speed while sacrificing a slight amount of accuracy.

RAQUIBA SULTANA and TETSURO NISHINO of The University of Electro-Communications, Japan presents their work "Fake News Detection System using BERT and Boosting Algorithm". This paper addresses the issue of false information spreading rapidly on social media and proposes an ensemble model based on transformers to identify false information, particularly related to Covid-19. The model utilizes a hybrid ensemble learning approach and achieves excellent accuracy (0.99) and F1 score (0.99). The Receiver Operating Characteristics (ROC) curve demonstrates a high true-positive rate and an AUC score of 0.99, indicating the effectiveness of the suggested model in identifying false information online.

INDRANIL ROY from Southeast Missouri State University, Cape Girardeau, NICK RAHIMI from the University of Southern Mississippi, Hattiesburg, BIDYUT GUPTA from Southern Illinois University, Carbondale, and NARAYAN DEBNATH from Eastern International University, Vietnam present their paper "Secured Communication in Generalized Non-DHT-based Pyramid Tree P2P Architecture". This paper discusses a structured P2P network with an interest-based system consisting of different clusters. While the network offers efficient data look-up protocols, it has the limitation of assuming that peers in a cluster can only have one resource type. The authors aim to overcome this restriction by generalizing the architecture and have made significant initial progress. They plan to modify existing data look-up protocols while maintaining low latency and consider security by using a public key-based approach. Preliminary results indicate that the use of public-private key pairs will be more efficient than a symmetric key-based approach.

NARAYAN DEBNATH from Eastern International University, Vietnam, CARLOS SALGADO, MARIO PERALTA, DANIEL RIESCO, LORENA BAIGORRIA, and GERMÁN MONTEJANO from Universidad Nacional de San Luis, Argentina presents their work “An evaluation Framework to ensure the quality of the conceptual Models of Business Processes in a Biodiesel Plant”. This work proposes improving the quality of business processes in a biodiesel plant by analyzing and studying the conceptual models of the processes. A framework is applied to measure the quality, providing metrics and indicators for evaluation. The framework facilitates the understanding, maintenance, and evolution of business processes, benefiting organizations by ensuring comprehension, reducing efforts for model changes, and supporting early evaluation of quality properties. The framework consists of two evaluation methods, one numerical and the other based on linguistic expressions, providing valuable results for different areas of the business.

Ajay Bandi, CATA 2023 Program Co-Chair
Mohammad Hossain, CATA 2023 Program Co-Chair
Ying Jin, CATA 2023 Conference Chair

IJCA Contributed Papers: As mentioned earlier, the second part of this issue comprises papers that were contributed to the International Journal of Computers and their Applications (IJCA). The topics and main contributions of the papers are briefly summarized below:

TOMAZ KOKOT from Alma Mater Europaea, Austria presents his work “Hybrid Remote Work Models in Project-Organized Small and Medium-Sized IT Companies”. According to this report, remote work has become increasingly popular in the ICT industry due to the COVID-19 pandemic, resulting in changes in workplace dynamics and employee perspectives. Remote work provides numerous benefits, including flexibility, a better work-life balance, and cost savings, as well as reducing employee turnover and boosting productivity. A survey of 100 employees from small and medium-sized IT companies showed positive results, with remote work enhancing motivation, satisfaction, and productivity. The ICT industry may need to make improvements to support remote work or adopt a hybrid model that combines the best features of remote work and traditional work methodologies. Although the future of ICT work is uncertain, it is expected to be a dynamic and adaptable industry.

SUCHETA SABLE and RAJESH KHERDE from DY Patil University, India presents their work “Prediction of water quality parameters using ARIMA & WPFM”. This paper utilizes machine learning methods, such as autoregression with mean average and Random Forest Regression, to forecast water data from the Godavari River in Maharashtra, India. The models are compared, and the ARIMA model is found to be reliable for predicting the subsequent six values with an RSME score of 0.5. Additionally, samples from the study sites are collected and tested in the lab to compare the predicted results with the actual values during the machine learning process.

RATTANAWADEE PANTHONG from the University of Phayao, Thailand presents his work "Hybrid SMOTE and Bootstrap Sampling for Imbalanced Classification in Elderly Health Condition Dataset". The study analyzes real-world data imbalances in elderly health condition datasets, using hybrid data level techniques like SMOTE and Bootstrap approaches. Four learning methods, deep learning, stacking algorithm, random forest, and gradient boosting tree, are applied to improve classification accuracy. The HySM_BT50% method achieved the highest correctness value at 90.11% using random forest as a classifier.

MOATAZ MOHAMMED, SALSABIL A. EL-REGAILY, and MOSTAFA M. AREF from Ain Shams University, Egypt present their work "End-To-End Open-Domain Question-Answering System: Baseline and Case Study using EIAD Dataset". This research focuses on the Open-Domain Question-Answering (ODQA) task in the field of Islamic religion. The IslamBot QA system is developed using deep learning-based retrieval-reader models, using data from the English Islamic Articles Database (EIAD). The model sets a new standard with Dense Passage Retriever models, achieving 78% R@100. The model generates novel results, with a 71.5% EM and 75.8% F1 score. However, the long-form open-domain type poses challenges in justification and input without context.

KALIM QURESHI, FATIMA YOUSEF, and PAUL MANUEL from Kuwait University, Kuwait present their work "Uml Framework: National Institute for Health and Care Excellence (Nice) Diabetes Guidelines Based Diabetes Information System". This paper addresses the need for an effective diabetes management system and proposes a Diabetes Management Information System (DMIS) based on the NICE guidelines. The DMIS is designed using the Unified Modeling Language (UML) and validated through traceability techniques and expert involvement. The results indicate that the proposed DMIS model is comprehensive and aligns with all the recommendations provided by NICE, offering a valuable tool for diabetes care and treatment.

BAGHDADI AMMAR AWNI ABBAS and MOHAMMED AL-MUKHTAR from the University of Baghdad, Iraq present their work "Building Computer-Based Test (CBT) using MATLAB: programming the essential types of questions". This paper proposes a MATLAB testing package for creating Computer-Based Tests (CBT) using six common question types in Learning Management Systems (LMS). The package uses the MATLAB App Designer tool and an Excel spreadsheet for exam information, student answers, and grades. Users can create unlimited exam questions and choose between real tests or training. Grading is a mixed operation between the user and the computer, with the program being static and 100% accurate for 200 users attempting 20 different exams.

As guest editors, we would like to express our deepest appreciation to the authors and the reviewers. We hope you will enjoy this issue of the IJCA. More information about ISCA society can be found at http://www.isca-hq.org.

Guest Editors:

Ajay Bandi, Northwest Missouri State University, USA

September 2023

Object Recognition for Autonomous Vehicles from Combined Color and LiDAR Data

Lian Kang and Pierre Payeur
University of Ottawa, Ottawa, Ontario, CANADA

Abstract

In recent years, autonomous driving vehicles have garnered substantial attention in both the commercial and scientific domains. A key challenge faced by these vehicles is the accurate detection and recognition of objects within complex real-world road environments, essential for their real-time decision-making capabilities. While color imaging has traditionally provided rich information, the utilization of LiDAR scanners presents advantages such as high-quality data collection under varying lighting conditions and the provision of precise spatial information with an extensive range. By combining data from color cameras and LiDAR scanners, the potential for object detection in autonomous driving is expanded, opening up new avenues for advancement. This paper introduces a novel 3D object detector that leverages a bird's-eye view map generated from a LiDAR point cloud along with RGB images as input data. It employs focal loss and Euler angle regression techniques to enhance object detection performance. Through ablation experiments, the achieved improvements are evaluated. Experimental results demonstrate the framerate and performance of the proposed 3D object detector, surpassing 46 frames per second and achieving an average precision exceeding 90%. Additionally, a more compact version of the detector is introduced, processing the same input data three times faster while maintaining reasonably high accuracy.

Key Words: Object recognition, deep learning, LiDAR, autonomous vehicles.

1 Introduction

The advent of artificial intelligence has propelled autonomous driving vehicles into the realm of possibility, garnering substantial investments from major players like Tesla and Waymo. Detecting and recognizing pedestrians, vehicles, and other objects on the road is a crucial task in autonomous driving systems, ensuring safe and informed driving decisions. As a result, there has been a significant focus on developing efficient object detection and recognition technologies.

Deep learning methods have revolutionized the field of object detection and recognition, demonstrating remarkable advancements. Recent research has emphasized the inclusion of depth information in detection models, surpassing the limitations of traditional 2D mapping approaches for autonomous driving. Light detection and ranging (LiDAR) scanners, unlike stereo cameras and active depth sensors, offer consistent and high-precision spatial information unaffected by ambient lighting conditions. Consequently, LiDAR scanners have gained popularity for 3D object detection in outdoor environments. However, while LiDAR scanners provide shape and location information of objects in the real world through point cloud data, they lack texture and color information. To fully exploit 3D location with color and texture information, the point cloud from LiDAR scanners often needs to be preprocessed and combined with RGB images.

This paper represents an extended version of [6] providing further insights into the methodology employed, along with conducting comprehensive experimental ablation studies and robustness validation. The aim is to delve deeper into the intricacies of the proposed approach and to evaluate its performance under various scenarios and conditions. By conducting these additional analyses, a more comprehensive understanding of the methodology's strengths, limitations, and overall effectiveness can be achieved. The key original contributions of this research include the integration of Euler angle regression into the DarkNet-53 convolutional neural network [14] to create a 3D object detector capable of classifying and localizing cars, pedestrians, and cyclists using LiDAR point clouds and RGB images from real-world road scenes. To optimize training and testing efficiency, the LiDAR point cloud is transformed into a bird's-eye view (BEV) map using coordinate system transformation and height thresholding. Furthermore, the proposed architecture incorporates a focal loss [10] and a generalized intersection over union (GIoU) loss [15] to address biased data and to enhance the model's performance. The solution is trained and evaluated on real-world data provided by the KITTI vision benchmark suite [4], demonstrating its object recognition capabilities.

2 Literature Review on 3D Object Detection from LiDAR Data

In recent years, significant advancements in artificial intelligence have been made in the field of 3D object detection and recognition for autonomous driving. To fully support 3D object detection, which involves an important component of localization in the environment, the utilization of not only traditional RGB or grayscale images but also depth information

* Email: lkang018@uottawa.ca, ppayeur@uottawa.ca.

is required as it provides the spatial coordinates of each pixel. Consequently, the need for larger and more complex training and testing datasets arises to build performing deep learning models, posing higher demands on data processing and computing capabilities.

Two-stage detectors, such as MV3D-Net [3] and AVOD [7], have gained prominence in this field. MV3D-Net incorporates both RGB images and LiDAR point clouds as input. It combines a 3D object proposal network and a region-based fusion network to efficiently generate 3D candidate boxes from bird's-eye view (BEV) maps and front-view perspectives derived from the point cloud. AVOD shares similarities with MV3D-Net but utilizes a feature pyramid network (FPN) instead of a VGG16 based network for feature extraction. FPN helps maintain feature map resolution and preserves both low-level and high-level information, leading to enhanced detection accuracy, particularly for small objects.

Single-stage detectors, such as PIXOR [23] and PointPillars [8], have also made notable contributions to 3D object detection. PIXOR discretizes the point cloud by equally spaced units and encodes reflectivity to create a regular representation. It employs a fully convolutional network (FCN) to estimate the position and heading angle of the target relative to the sensor. PIXOR achieves a high detection framerate of over 28 frames per second (FPS). On the other hand, PointPillars directly aggregates points falling into each grid, forming ‘Pillars’. The learned feature vector is then mapped back to grid coordinates, resembling an image-like representation.

In addition to these methodologies, researchers have integrated the Transformer architecture [19] into 3D detection algorithms. VoTr [11] employs a voxel-based Transformer as a 3D backbone network for object detection from point cloud data. It utilizes self-attention mechanisms to establish long-range relationships between voxels. To improve attention span without excessive computational overhead, VoTr introduces local attention, dilated attention, and fast voxel query. The versatility of the Transformer architecture is demonstrated through variants such as VoTr-SSD and VoTr-TSD [11], which leverage SSD and R-CNN backbones, respectively.

In summary, notable progress has been made in 3D object detection and recognition for autonomous driving applications through various approaches. Two-stage detectors, such as MV3D-Net and AVOD, offer efficient fusion of RGB images and LiDAR point clouds, while single-stage detectors like PIXOR and PointPillars provide rapid detection capabilities. Additionally, the incorporation of the Transformer architecture, exemplified by VoTr, shows promise in capturing long-range dependencies in point cloud data.

3 Point Cloud Preprocessing and Registration with RGB Data

A 3D point cloud is the default representation of the data collected by a LiDAR scanner. The LiDAR scanner used to collect point clouds for the KITTI dataset [4] and considered in this research is the Velodyne’s HDL-64E. It is a 64-channel multi-beam mechanical LiDAR that continuously rotates the head to achieve dynamic 3D scanning. It covers a 360° horizontal and 26.9° vertical field of view [4]. Although the data provided by LiDAR reports an accurate 3D location, it does not contain color information. Therefore, in this work, both the color images provided by a collocated RGB camera, and the 3D point clouds provided by the LiDAR are used.

There are two major ways to process a 3D point cloud: one is directly processing a 3D matrix, while an alternative approach involves projecting the 3D map to its corresponding 2D representation. In recent work using LiDAR point clouds for object detection and recognition [5, 9, 26], researchers directly train the detector on the 3D point clouds [18] with the consequence that convolution operations are more time and memory intensive compared to when information is encoded in 2D. Alternatively, some models like MV3D-Net [3] opt for converting the point cloud to a front view and a BEV map, which is more efficient than processing the 3D point cloud directly, while not lowering the accuracy.

The detector proposed in this work uses BEV maps that are initially encoded as 2-channel bidimensional grids where each pixel of a 2D map contains respectively a ‘cumulated height’ parameter associated with 3D points mapped to the pixel, and a ‘cumulated intensity’ estimate provided by the LiDAR sensor along with every 3D point measurement. The BEV maps converted from point cloud data collected by the LiDAR scanner are further combined with registered front forward view images collected by the RGB camera as inputs. This section details how the LiDAR point clouds are preprocessed and converted to the BEV maps, as shown in Figure 1.

3.1 Data Registration

The BEV map is a graphical representation of the point clouds from a bird’s-eye view. It is obtained by projecting the discrete LiDAR point cloud on a plane perpendicular to the height direction. Therefore, a BEV map forms an equivalent representation of the 3D location information contained in the LiDAR point cloud, hence it is more efficient when a large amount of data is being processed. The front forward view and color information is provided from corresponding RGB images. Therefore, all the required information can be obtained by combining the BEV maps converted from LiDAR point clouds and the RGB images.

The 3D point cloud and RGB images are obtained from different sensors and must be registered before the two sets of data are used together as input. With the help of the calibration matrices from the dataset, the registration of the LiDAR point clouds and RGB images is performed using matrix transformation to align LiDAR coordinate axes and origin to the RGB coordinates.

3.2 Mapping 3D Points within Region of Interest to 2D Pixels

In the dataset considered, the point cloud of each scene represents approximately 1.9 MB, which requires significant memory space, highly increases computation for both training and testing, and reduces detection efficiency. Therefore,

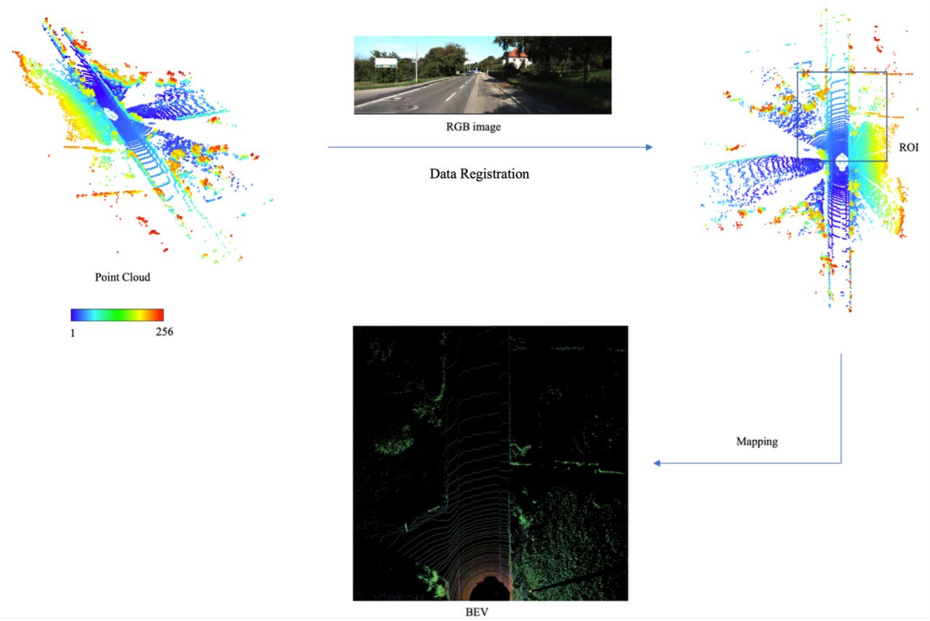


Figure 1: 3D point cloud preprocessing and registration with RGB image.

detection is focused over a pre-selected region of interest (ROI) in the point cloud. To balance the model's efficiency while covering all the annotated target objects in the corresponding RGB image, the ROI of the point cloud is manually set as a rectangular area that spans 40m on either side of the LiDAR scanner, and 80m in front of it. The point cloud data collected from the LiDAR scanner are 3D points with real (x, y, z) values that carry depth information. The registered points within the ROI carrying real number values are then mapped into integer values (u, v) that represent the pixel location on the discretized bird's-eye view (BEV) map.

3.3 Recording the Height and Intensity Information in the BEV Map

Once the coordinates (u, v) that represent each pixel in the BEV map are obtained, the height information represented by the values on the Z-axis and the intensity information represented by a 4th parameter contained in the source point cloud matrix are extracted from the 3D point cloud to be encoded in the corresponding BEV map.

Inspired by the representation adopted in PIXOR [23], a vertical ROI on the Z-axis is selected to support a height thresholding operation that is applied to preserve only data from the point cloud that are within the selected height of the ROI. Next, the height coordinates (on Z-axis) within the ROI are rescaled to the]0, 255[range, and the height coordinates exceeding the ROI are forced to 0 or 255. Finally, the height values of the points in the point cloud that are mapped into the same 2D pixel position on the BEV map are cumulated and recorded to the "cumulated height" channel of the BEV map. Compared with using the maximum height of each pixel position [1], the cumulation method appears to be less affected by changes in the elevation of the objects due to the characteristics of the environment, such as the slope of the road. Unlike MV3D-Net [3] that manually selects multiple ranges on the Z-axis and accumulate the values of the points within the ranges to generate multiple height channels for each 2D point in the BEV map, the proposed method performs a single height thresholding operation to create one height channel. This contributes to improve the efficiency of the detection process.

Similarly, the intensity information already contained in the point cloud as the 4th value for each 3D point is extracted. For all 3D points contained within the selected height ROI and falling within the same 2D pixel position on the BEV map, these intensity values are accumulated and recorded to the 'cumulated intensity' channel of the BEV map. The resulting preprocessed point cloud data leads to a 2-channel bidimensional BEV map which is used as part of the input for the proposed 3D object detector along with the corresponding 3-channel RGB image input.

4 3D Object Detector Architecture

The proposed method for 3D object detection combines the BEV map generated from a LiDAR point cloud and the associated RGB image information to form a single 5-channel (height, intensity, R, G, B) input for every 2D pixel coordinate in the BEV map, as depicted in Figure 2a. The BEV map part of the input represents the bird's-eye view over the detection range, with a cumulated height channel and cumulated intensity channel, as detailed in Section 3. The RGB image is subsampled and padded with $[R, G, B] = [128, 128, 128]$ to match the size of the BEV map, as shown in Figure 2a. It forms the RGB component of the input, which represents the front forward view as found in autonomous driving, with three different color channels (R, G, B). Doing so preserves both the 3D information collected by the LiDAR scanner in the distribution of feature points in the 2D BEV map and the color information collected by the RGB camera.

The output of the proposed model represents the detection and recognition confidence over three object classes (car, pedestrian, cyclist), with prediction matrices at three different scales. The prediction matrices are used to draw bounding boxes (B-Box) around detected objects and to label them with their respective classification. Given the importance of making fast decisions in autonomous driving, any improvement in object detection speed while maximizing object detection and classification accuracy is prioritized. For this reason, a single-stage detection model is proposed in this paper.

As shown in Figure 2, the backbone of the proposed object detection model is based on DarkNet-53 [14], modified to include the preprocessing stage of the BEV maps and the RGB images described in Section 3. The detection head is based on the YOLOv3 anchor regression method, modified by adding BEV variables and rotation angle regression. For the latter, the rotation angle encoding uses the Euler representation, as inspired by complex YOLO [17].

4.1 Detection Head with Euler Angle Regression

Through the backbone convolutional neural networks and the FPN layers feature maps at three different scales are extracted from the input. As shown in Figure 2d, a detection head is used to generate the detection and recognition results based on these feature maps.

Within the detection head, each feature map is divided into grid cells. For each grid cell there are three anchors at different scales. These anchors represent priors for bounding boxes (B-Box). They are equivalent to a reference frame for the predicted B-Box. Based on this reference, the predicted B-Box generated by the detection head only needs to be fine-tuned with respect to the corresponding anchor. As a result, for every anchor, there is a prediction matrix that contains the parameters used for regression during training and the detection result. The output prediction matrix of the proposed detection head contains the B-Box prediction matrix for both the RGB front forward view and the BEV, a confidence score, and classification scores over the 3 object classes considered.

To adapt to the different perspectives of the predicted input, based on Yolov3 [14], the B-Box prediction matrix for the proposed detection head is modified. It is divided into two parts: one for the front forward view, another one for the BEV, as shown in Figure 3. The prediction matrix contains 14 parameters (i.e., N = 14 in Figure 2d). The confidence score p_0 indicates the confidence that the predicted B-Box contains an object. If this predicted B-Box corresponds to the background, then this value should be 0. The classification scores p_1, p_2, p_3 represent the probability that the category of the predicted B-Box corresponds to 'car', 'pedestrian', or 'cyclist' respectively. For the final output, only the B-Box with p_0 higher than a detection threshold will be kept and the classification shows $\max(p_1, p_2, p_3)$.

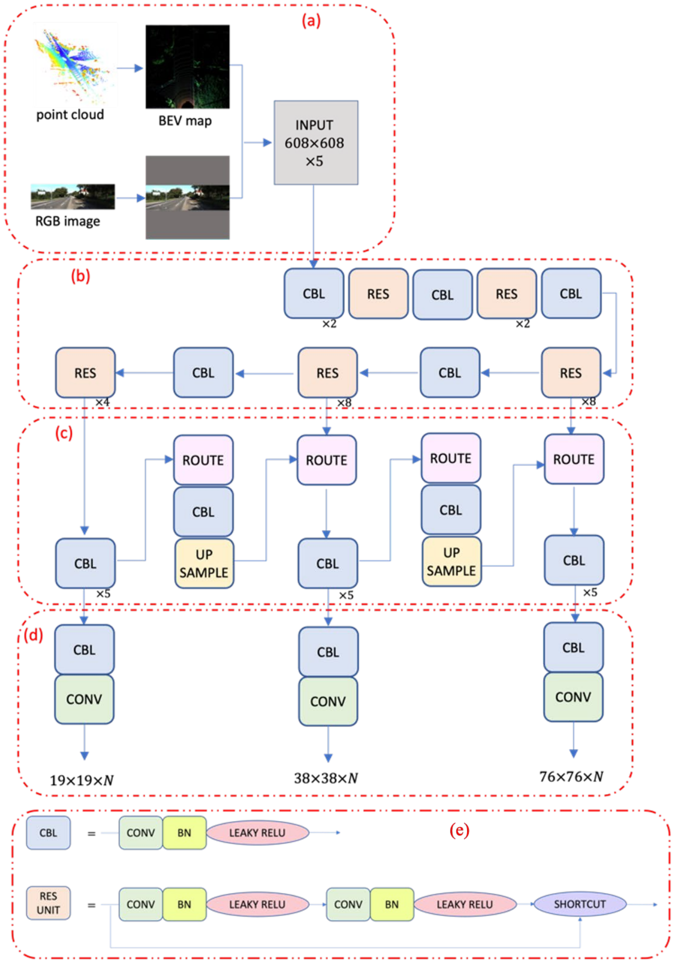
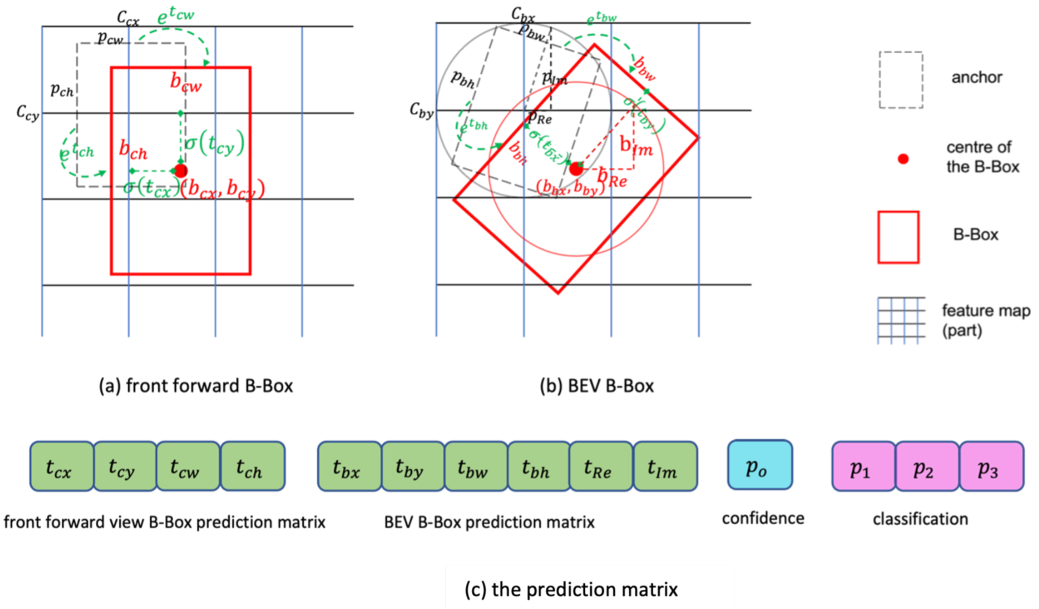


Figure 2: Proposed 3D object detection model architecture: (a) preprocessing stage to convert a 3D point cloud and corresponding remapped RGB image into a 5-channel 2D BEV map; (b) backbone of the proposed model (DarkNet-53); (c) feature pyramid network (FPN); and (d) detection head with outputs prediction matrices at three different scales; with (e) details of the respective structure of CBL (top) and res unit (bottom)

Considering that objects of primary interest in the context of autonomous driving, such as cars, pedestrians, and cyclists can generally be assumed to stand or move on the ground, the front forward view bounding box (B-Box) surrounding a detected object on the RGB image can be represented by 4 parameters. These are the center coordinates, (t_{cx}, t_{cy}), and the width and height, (t_{cw}, t_{ch}), of the B-Box as shown in Figure 3a.

Unlike the B-Box on RGB images, the BEV B-Box might not be parallel to the BEV map's coordinate axes, as exemplified in Figure 3b. To predict the relative rotation of the B-Box, as inspired by complex YOLO [17], a Euler representation of the rotation angle is added to the prediction matrix in the proposed detection model. Hence, the BEV prediction matrix obtained from the regression of the proposed model contains 6 variables.



(a) front forward B-Box

(b) BEV B-Box

(c) the prediction matrix

Figure 3: Converting the prediction matrix into B-Box: (a) front forward view B-Box; (b) BEV B-Box; and (c) prediction matrix of the detector with 14 parameters

Aside from the offsets of the B-Box center coordinate, (t_{bx}, t_{by}), and its width and height (t_{bw}, t_{bh}), the Euler representation of the rotation angle of the B-Box, (t_{Im}, t_{Re}), is also encoded in the 6 parameters.

4.2 Loss Functions

The loss function used in the proposed detector model consists of a combination of a classification loss, a B-Box regression loss, and a confidence loss. Compared to YOLOv3 [14], the regression loss uses Generalized Intersection over Union (GIoU) [15] instead of mean square error (MSE). To further optimize the performance of the detector, the Euler angle is added to the B-Box regression loss. For the classification loss, a focal loss [10] is added to address the imbalance problem observed in the KITTI vision benchmark suite [4] training dataset for the considered three classes.

4.2.1 Regression Loss. To match with the Euler angle regression network, a combination of GIoU [15] and Euler angle regression is used for the B-Box regression. The GIoU of the predicted B-Box and ground truth B-Box is computed as:

$$GIoU = \frac{I}{B^g \cup B^p} - \frac{A^c - (B^g \cup B^p)}{A^c} \tag{1}$$

Where B^g, B^p are the ground truth B-Box and the predicted B-Box respectively. I is the intersection of the ground truth and predicted B-Boxes, and $B^g \cup B^p$ is the union of the two B-Boxes. A^c represents the smallest convex shape that encloses both B^g and B^p. $A^c - (B^g \cup B^p)$ represents the area that is inside A^c but outside $(B^g \cup B^p)$. The GIoU loss for the front forward RGB view is represented by:

$$L_{GIoU} = 1 - GIoU \tag{2}$$

Since the B-Box prediction matrix contains 4 variables for the front forward view and 6 variables for the BEV, the B-Box regression loss is divided into two parts: GIoU loss (L_{GIoU}) for the front forward view prediction matrix, and an Euler defined GIoU loss (L^E_{GIoU}) for the BEV prediction matrix which is also computed with Equation (2) but on the BEV view.

4.2.2 Classification Loss. In YOLOv3 [14], cross entropy loss [25] is used for classification. In ideal circumstances, a non-biased training dataset helps the model learn the features for multi-class object detection and recognition, and cross entropy loss would be suitable. However, among the three classes considered in this research (car, pedestrian, cyclist), the training data provided by KITTI [4] contains 82% of the total objects in the class ‘car’, 13% in the ‘pedestrian’ class, and less than 5% in the ‘cyclist’ class. This represents a significant bias toward the ‘car’ category which must be addressed to achieve fair comparative results. Inspired by [10], a focal loss is used to substitute the cross entropy loss. This strategy represents a first usage of focal loss on BEV maps to the best of our knowledge. The classification loss is defined as:

$$L^F_{cla} = \sum_{i=0}^{S^2} \sum_{j=0}^{B} \mathrm{I}_{i,j}^{obj} \sum_{c \in \text{classes}} [\hat{p}_c (1 - p_c)^\gamma \log(p_c) + (1 - \hat{p}_c) \hat{p}_c{}^\gamma \log(1 - p_c)] \tag{3}$$

$$I_{i,j}^{obj} = \begin{cases} 1, & \text{if object} \in \text{the } j^{\text{th}} \text{ anchor} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Where γ is the relaxation parameter. The higher value of γ, the more 'focus' will be given to misclassified examples, and the less loss will be propagated from examples. S^2 represents the number of grid cells, which is equal to the size of the feature map. In the proposed model with three different scales, S^2 has sizes 19×19, 38×38, 76×76, respectively. B represents the B-Box. $\mathrm{I}_{i,j}^{obj}$ is a binary value that indicates whether the j^{th} B-Box of the i^{th} grid cell's GIoU value is larger than the GIoU threshold. p_c and $\hat{p}_c$ are the ground truth and the prediction classification score for class c.

4.2.3 Confidence Loss. The confidence loss is used to measure the objectiveness of the B-Box. The proposed model uses focal loss as its confidence loss, defined as follows:

$$L_{con} = \sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{C}_i(1-C_i)^{\gamma}\log(C_i) + \left(1-\hat{C}_i\right)\hat{C}_i^{\ \gamma}\log(1-C_i)\right] + \lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{noobj}\left[\hat{C}_i(1-C_i)^{\gamma}\log(C_i) + \left(1-\hat{C}_i\right)\hat{C}_i^{\ \gamma}\log(1-C_i)\right] \tag{5}$$

$$I_{i,j}^{obj} = \begin{cases} 1, & \text{if object} \in \text{the } j^{\text{th}} \text{ anchor} \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

$$I_{i,j}^{noobj} = \begin{cases} 1, & \text{if the } j^{\text{th}} \text{ anchor is background} \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

$$\hat{C}_i = \hat{p}_i(c) \times (GIoU + GIoU_E) \tag{8}$$

$$C_i = \begin{cases} 1, & \text{if object} \in \text{the } j^{\text{th}} \text{ anchor} \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

If an object is detected in the B-Box, the confidence loss is $\sum_{i=0}^{S^2}\sum_{j=0}^{B} I_{ij}^{obj}\left[\hat{C}_i(1-C_i)^{\gamma}\log(C_i) + \left(1-\hat{C}_i\right)\hat{C}_i^{\ \gamma}\log(1-C_i)\right]$. $\hat{C}_i$ is the confidence score of the j^{th} prediction B-box in i^{th} grid cell, and C_i is the ground truth, that is whether the B-Box contains an object.

In realistic scenarios, most bounding boxes do not contain any object. This causes an imbalance problem where the background or negative samples are more frequently detected by the model than the objects of some positive samples. To alleviate this issue, the confidence loss is weighted down by a factor λ_{noobj}, which intervenes when no object is detected in the box (detected background only). In such a case, the confidence loss is $\lambda_{noobj}\sum_{i=0}^{S^2}\sum_{j=0}^{B} \mathrm{I}_{ij}^{noobj}\left[\hat{C}_i(1-C_i)^{\gamma}\log(C_i) + \left(1-\hat{C}_i\right)\hat{C}_i^{\ \gamma}\log(1-C_i)\right]$, where I_{ij}^{noobj} is the complement of the binary value I_{ij}^{obj}, and λ_{noobj} weighs the loss down.

In summary, the loss function of the proposed 3D object detector combines the two GIoU regression losses (L_{GIoU} applied on the front forward view and the Euler angle loss L_{GIoU}^{E} applied on the BEV view), the focal classification loss (L_{cla}^{F}), and the confidence loss (L_{con}). The combination of all components leads to the general loss function:

$$L = \alpha_1 L_{cla}^{F} + \alpha_2 L_{GIoU} + \alpha_3 L_{GIoU}^{E} + \alpha_4 L_{con} \tag{10}$$

Where α_1, α_2, α_3, α_4 are the weights for each component of the loss function, which are empirically determined based on experimental results.

5 Experimental Results

5.1 Performance Evaluation

The performance of the proposed architecture is compared with other popular 3D object detectors that also use the KITTI LiDAR data as input. Table 1 summarizes the mean average precision (mAP) performance and framerate achieved while considering three difficulty levels for the object detection process, as defined in the KITTI evaluation metrics. The category 'easy' represents cases where low objects occlusion occurs and with objects' B-Box height reaching over 40 pixels in the front forward RGB image. 'Moderate' cases involve objects that are at least partially visible and taller than 25 pixels. Finally, 'hard' cases correspond to significantly occluded objects that are difficult to observe in images.

In terms of the detector framerate, the experimental evaluation demonstrates that the proposed model reaches 46.4 frames per second (FPS) in an implementation using a single NVIDIA Tesla V100 GPU. This is more than 3 times faster than the transformer-based detector VoTr-SSD, 3 to 4 times faster than two-stage detectors with similar detection accuracy, and faster than PointPillars by over 4 FPS with similar or exceeding accuracy.

The proposed detector also demonstrates superior precision performance compared to other models listed in Table 1, including the recently introduced transformer architectures. In this case, it outperforms the dominant two-stage F-ConvNet detector by a significant margin of over 3%. Performance gains are particularly visible in cases categorized as 'easy', as illustrated in the corresponding results shown in Figure 4. For visualization purposes, 3D B-Boxes are plotted around the detected objects over the testing RGB images (top part of results), and the corresponding 2D BEV B-Boxes are plotted over the BEV maps (lower part of results). The color coding of the B-Boxes represents the class of the detected objects: yellow for 'car', red for 'pedestrian', and blue for 'cyclist'.

Compared with other detectors that use both the LiDAR point cloud and RGB images, such as MV3D-Net [3], AVOD [7], PIXOR [23], MMF [9], F-PointNet [12] and F-ConvNet [20], the proposed detector demonstrates higher mAP on 'moderate' and 'hard' samples. As shown in Table 1, some models that use only a LiDAR point cloud as input reach slightly higher mAP on 'hard' cases compared with models that combine LiDAR point clouds with RGB images as input. Sample experimental results achieved with the proposed LiDAR+RGB single-stage detector are presented for 'moderate' and 'hard' cases in Figures 5 and 6, respectively.

Table 1: Comparison of performance among 3D object detectors using the KITTI LiDAR dataset

	Method	Data	Framerate (FPS)	mAP (%)		
				Easy	Moderate	Hard
Transformer	VoTr-SSD [11]	LiDAR	14.7	87.86	78.27	76.93
	VoTr-TSD [11]	LiDAR	7.2	89.04	84.04	78.68
Two Stages	MV3D-Net [3]	LiDAR + RGB	2.7	86.49	78.98	72.23
	AVOD [7]	LiDAR + RGB	10.0	89.74	84.81	78.12
	F-PointNet [12]	LiDAR + RGB	5.7	91.16	84.61	74.77
	F-ConvNet [20]	LiDAR + RGB	1.9	**91.44**	85.84	76.11
	Fast Point R-CNN [2]	LiDAR	15.3	90.87	**87.71**	80.51
	MMF [9]	LiDAR + RGB	12.2	86.81	76.75	68.41
	STD [24]	LiDAR	10.0	89.93	86.20	79.42
Single Stage	VoxelNet [26]	LiDAR	4.2	87.95	78.39	71.29
	SECOND [22]	LiDAR	19.7	89.33	82.87	78.51
	PointPillars [8]	LiDAR	**41.9**	90.07	86.56	**82.81**
	SA-SSD [5]	LiDAR	24.4	88.75	79.79	74.16
	PIXOR [23]	LiDAR + RGB	28.6	86.78	80.75	76.77
	Proposed detector	***LiDAR + RGB***	***46.4***	***94.71***	***87.33***	***81.52***

Globally, when examining performance over all classes and independently from the 'easy', 'moderate', or 'hard' categorization of sample test cases, the mAP over all object classes reaches 90.26%, with the average precision (AP) for each specific class corresponding to 97.94% for 'car', 82.72% for 'pedestrian', and 90.13% for 'cyclist' respectively. Statistical details about the performance achieved are detailed in Table 2 when considering 1500 pairs of the LiDAR point cloud and the corresponding RGB images including all three classes of objects considered.

5.2 Ablation Studies with Different Loss Functions

The proposed detector merges Euler angle regression to the DarkNet-53 backbone to improve the detection accuracy and uses Euler angle regression loss and GIoU loss to optimize the training. To reduce the bias caused by the imbalance in the number of samples in each class of the dataset, focal loss [10] is also used as the classification loss and the confidence loss. To evaluate if these methods improve the performance of the proposed model, ablation studies are designed to test the performance of different regression loss and to verify that the proposed Euler angle regression does contribute to increase the detection accuracy of the proposed model. The results of the ablation experiments are listed in Table 3 where various combinations of loss functions are considered as the components of Equation (10).

The ablation studies with different regression loss show that the consideration of Euler angle regression increases the detection accuracy of the proposed object detector. Compared to MSE, GIoU loss also shows better performance on the regression of the proposed model. Finally, when comparing with the use of cross entropy as the classification loss, focal loss significantly increases the average precision (AP) for the 'pedestrian' and 'cyclist' classes although it slightly decreases AP for 'car'. Hence, it is concluded that focal loss does contribute to better balance the AP over all considered classes.

5.3 Masked BEV Map

In another set of experiments conducted as part of this research, a mask in the form of an empty rectangle was superimposed over the bird's-eye view map to selectively exclude certain information from the detector's input. This investigation aimed to determine whether the proposed model can handle damaged data. The experiments demonstrate that the proposed detector remains functional even when presented with damaged data, with only a marginal decrease in mean average precision (mAP) of less than 5%. Our observations revealed that in most cases if a target object is completely hidden or missing in the BEV map data, the proposed detector fails to detect the object reliably, as illustrated in Figure 7. This indicates that the detector cannot operate with reasonable accuracy solely based on the RGB image input. Conversely, if a randomly positioned mask partially occludes an object, as depicted in Figure 8, it generally does not significantly affect the detection and recognition outcome, as long as partial information about the target object remains available in the BEV map. Finally, when the mask covers a background area without occluding any target object of interest, as shown in Figure 9, the detection result remains unaffected.

6 Mini 3D Object Detector

As part of the continuous development of deep convolutional neural networks and aiming at always pursuing higher accuracy,

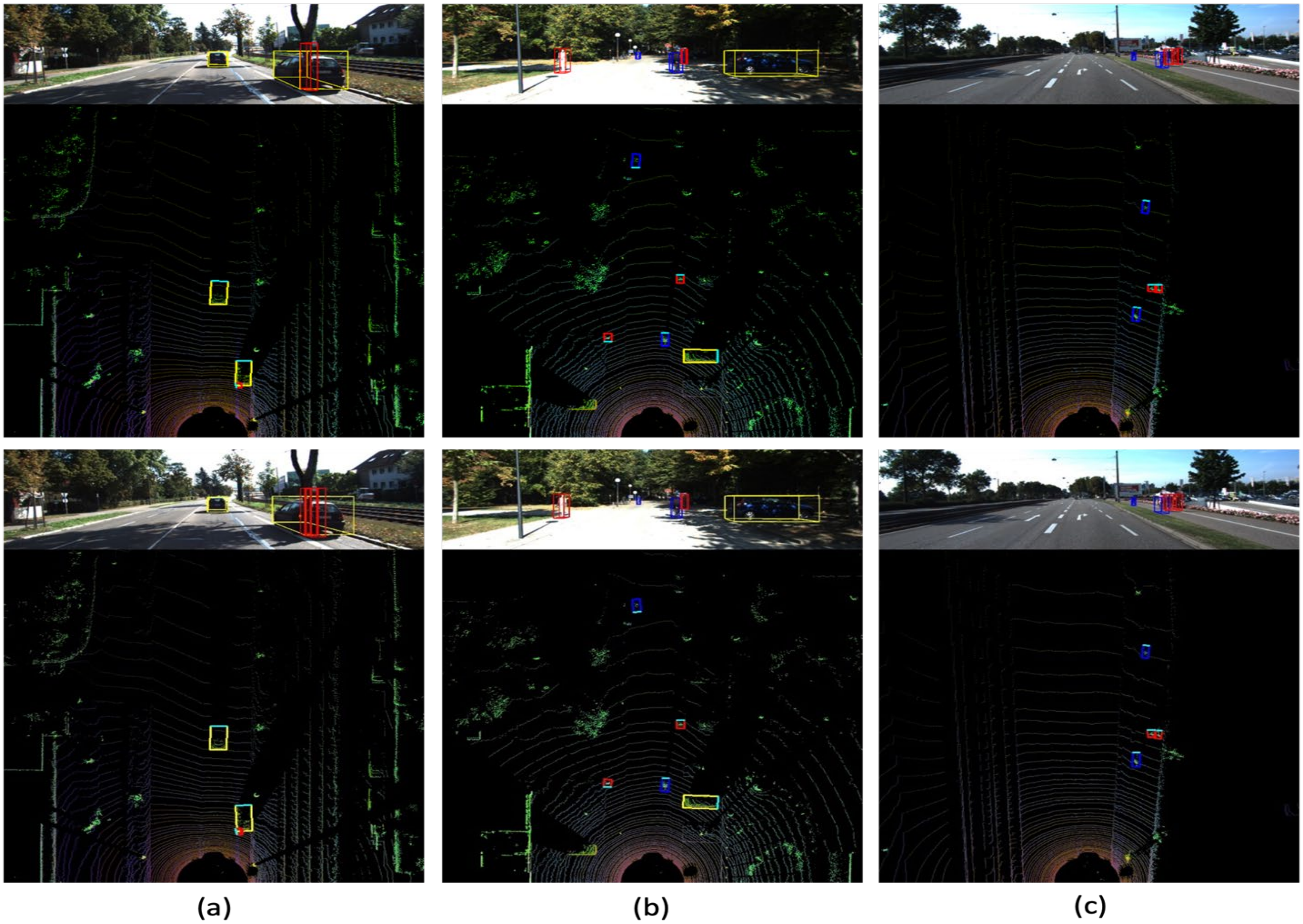


Figure 4: Samples of 'easy' scenes in the KITTI testing dataset. In each case (a,b,c), the upper row shows ground truth bounding boxes, and the bottom row shows detection results achieved with the proposed detector [yellow = car, red = pedestrian, blue = cyclist]

researchers are motivated to propose various strategies to increase object detection framerate, especially in speed sensitive context such as autonomous vehicles navigation. Ideally, a detector should be compact from both the memory requirement and amount of calculation perspectives, mainly because of the availability of limited hardware resources that can be embedded on autonomous platforms. These constraints motivate the development of deep convolutional neural architectures well adapted for widespread deployment on embedded devices. Therefore, although the framerate of the original detector introduced in Section 4 reaches up to 46.4 FPS, we still wish to explore the design of a lightweight network that involves fewer feature matrices to perform even faster on the same 3D object detection and recognition tasks.

6.1 Mini Detector Architecture

The proposed mini detector merges the structure of tiny-YOLO [16] and the proposed detector from Section 4 to generate feature maps at 2 different scales, as shown in Figure 10. The main difference of the mini detector compared to its full-size version is the backbone and FPN. The mini detector uses a DarkNet-19 [13] based backbone, modified to adapt to the input of the BEV map and corresponding RGB image. Compared to the 53-layers backbone of the original detector, the mini detector's backbone only has 19 layers, that is about 1/3 the depth. Moreover, with the mini detector implementation, only two different scales of feature maps are generated and passed to the detection head, compared to three in the initial version, to further reduce the calculation load and minimize the depth of the mini detector model. The detection head then converts the reduced size feature maps into prediction result. Otherwise, the detection head uses the same design as in the proposed full-size model and the same combined loss function, Equation (10), for training.

Being more compact, the mini detector's framerate can reach

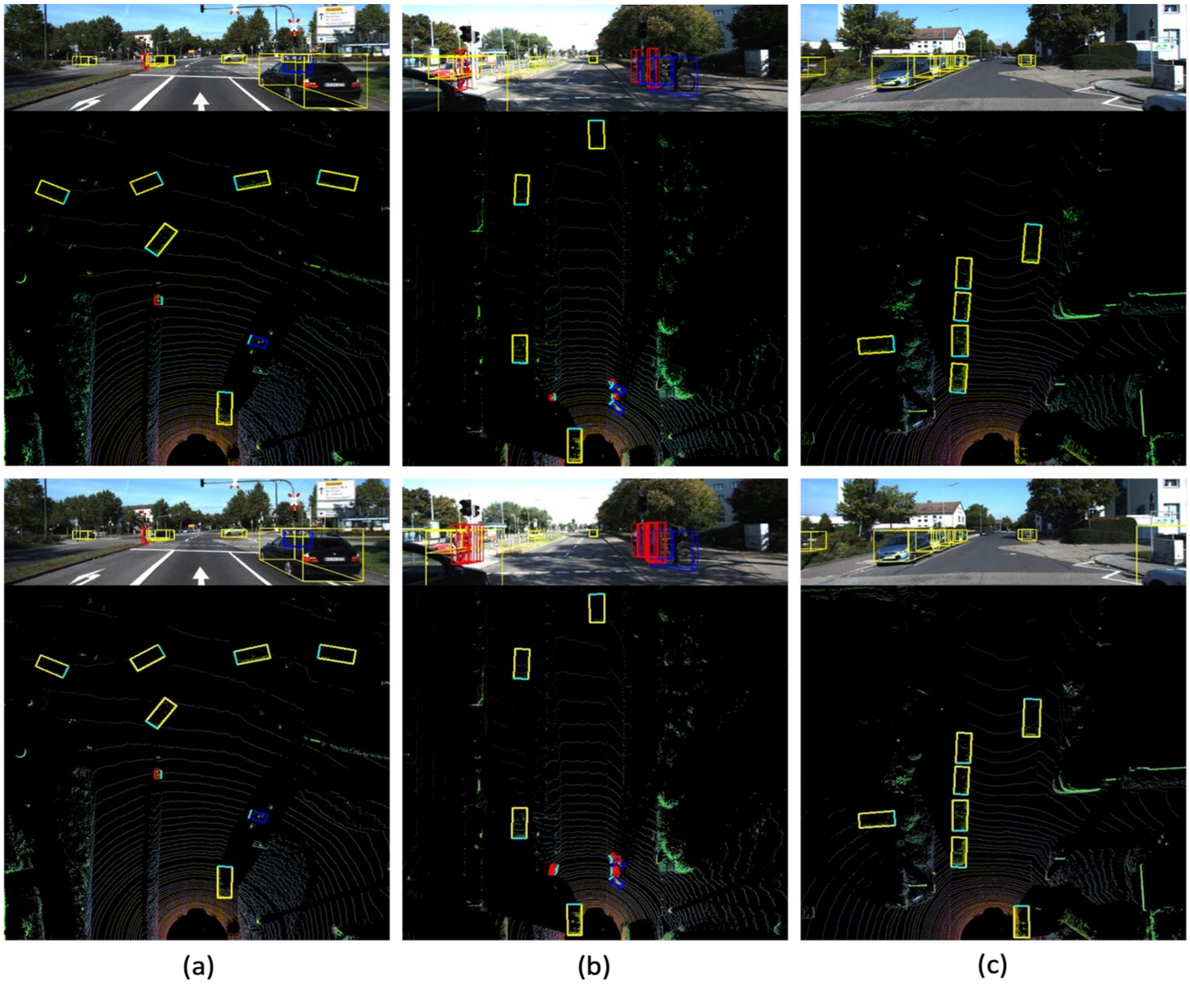


Figure 5: Samples of 'moderate' scenes in the KITTI testing dataset. In each case (a,b,c), the upper row shows ground truth bounding boxes, and the bottom row shows detection results achieved with the proposed detector [yellow = car, red = pedestrian, blue = cyclist]

up to over 3 times that of the original full-size detector. Therefore, it is better suited for real-time autonomous driving applications running on mobile devices, while offering a satisfactory compromise on detection performance.

6.2 Experimental Results with the Mini Detector

For a fair comparison of performance with the two proposed detectors, the mini detector is trained and tested on the same software and hardware environment as the original full-size detector presented in Section 4. Moreover, the training and testing dataset remains the same. Table 4 presents the detection and recognition results of both the mini detector and the full-size detector on 1500 pairs of the LiDAR point cloud and the corresponding RGB images.

As shown in Table 4, the mini detector achieves a good performance on detecting 'cars' with AP higher than 0.97, but lower AP is observed on detecting the 'pedestrian' and 'cyclist' targets. This is explained by the fact that the training dataset used for both proposed models is imbalanced, with less than 20% of positive samples exemplifying the pedestrian and cyclist classes. Although FPN and focal loss [10] are used to reduce the impact of the data imbalance, with fewer layers and less features extracted in the mini detector model, the testing performance is more severely impacted by the data imbalance than with the full-size detector. Figures 11 and 12 visually compare the performance against ground truth labels of both versions of the detector by displaying the predicted bounding boxes over the front forward RGB image and corresponding BEV map for detected objects belonging to the three considered classes.

Conversely, the framerate of the mini detector reaches up to 158.97 frames per second, which is 3.4 times faster than the full-size detector when testing in the same environment, while a 7.5% reduction of the mAP is observed overall on all three combined classes. Comparing with alternative compact implementations of objects detectors, as shown in Table 5, the proposed mini detector exhibits relatively high detection

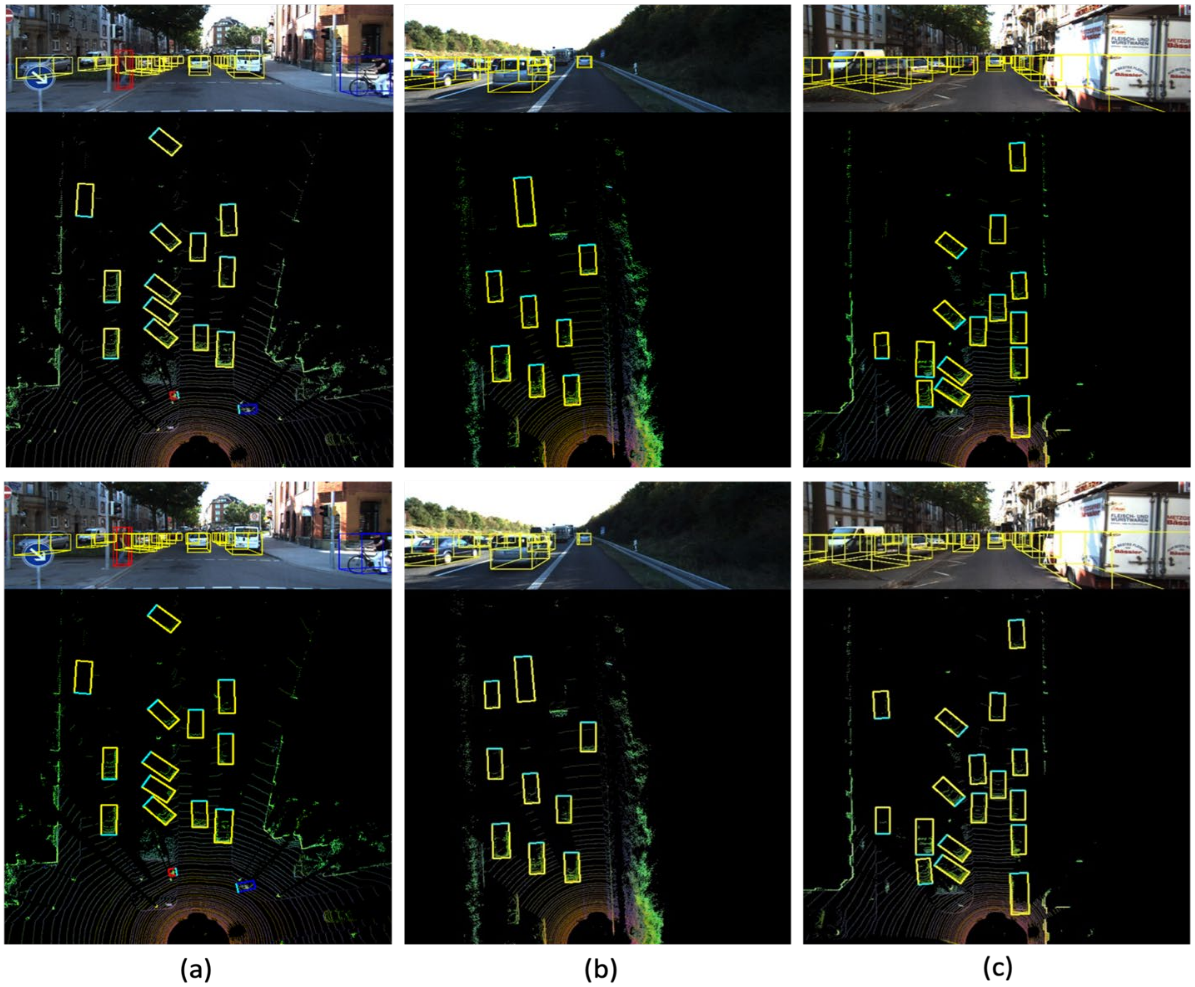


Figure 6: Samples of ‘hard’ scenes in the KITTI testing dataset. In each case (a,b,c), the upper row shows ground truth bounding boxes, and the bottom row shows detection results achieved with the proposed detector [yellow = car, red = pedestrian, blue = cyclist]

Table 2: Object detection performance on the KITTI LiDAR dataset

Class	Precision	Recall	AP	F1	mAP (%)	Framerate (FPS)
Car	90.65	98.68	97.94	94.50	**90.26**	**46.4**
Pedestrian	63.89	93.17	82.72	75.80		
Cyclist	79.51	95.24	90.13	86.67		

Table 3: Effect of different regression loss on the detection results

Classification Loss	Regression Loss	Confidence Loss	AP			mAP (%)
			Car	Pedestrian	Cyclist	
focal	MSE	focal	95.11	53.93	62.58	70.56
focal	GIoU	focal	96.78	64.61	83.83	81.74
focal	MSE + Euler	focal	96.88	78.48	90.96	88.77
focal	GIoU + Euler	focal	97.94	**82.72**	**90.13**	**90.26**
cross entropy	GIoU + Euler	focal	**98.03**	79.91	88.35	88.76

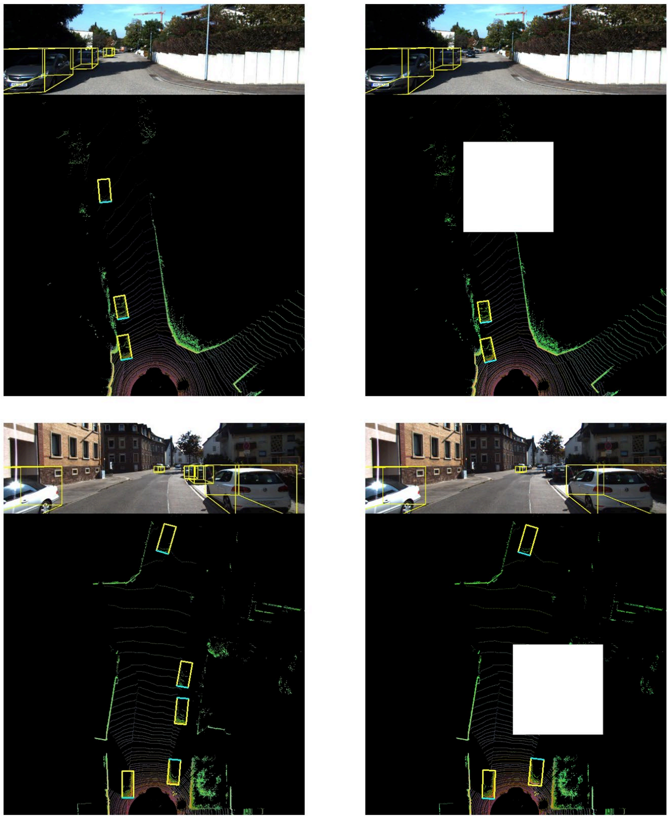


Figure 7: Objects detected on **s**ample test cases with fully masked objects in the BEV map (left: without mask; right: with occluding mask)

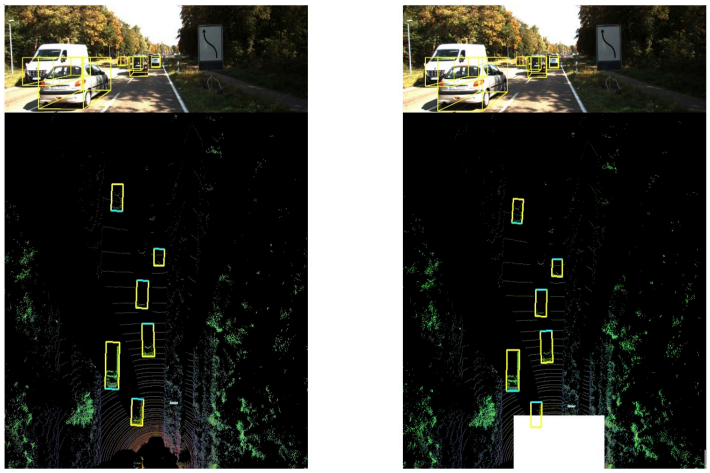


Figure 8: Objects detected on sample test case with partially masked object (left: without mask; right: with occluding mask)

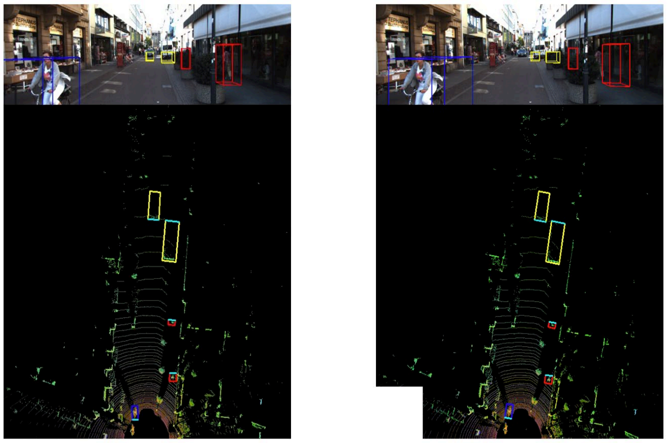


Figure 9: Objects detected on sample test case with mask over the background only (left: without mask; right: with occluding mask)

accuracy when compared to tiny YOLO [16] and tiny SSD [21], which are designed for 2D image-based only object detection. While expanding the architecture to benefit from 3D information issued from LiDAR data, the proposed mini detector remains competitive with the performance of similar scale detectors reported in the literature. The mini detector also achieves significantly higher detection framerate.

7 Conclusion and Future Work

This paper proposes two original formulations for 3D object detectors that leverage 3-dimensional location information from a LiDAR point cloud in combination with RGB images. With the objective to optimize the computation and memory usage of the detection models, a preprocessing step is performed to convert the point cloud data into a bird's-eye view (BEV) map. The latter emphasizes the height range of interest through height thresholding, while intensity values from the LiDAR sensor are accumulated and recorded on the corresponding pixel positions. When combined with color information from a registered frontal view RGB image, the process leads to a 5-dimensional BEV map that serves as input to the detectors.

The design of the proposed full-size detector model combines the GIoU loss and DarkNet-53 architecture from the YOLOv3 single-stage detector. Additionally, Euler angle orientation is incorporated into the detection head and an original formulation for a combined loss function is introduced. Experimental results reveal that the integration of Euler angle regression and GIoU losses enhances the performance of the proposed detector compared to the original YOLOv3 model, which utilizes MSE regression loss.

To address the bias in detection results caused by imbalanced training data, focal loss is employed as the classification loss. Ablation studies demonstrate that focal loss partially compensates for data imbalance in the proposed models and improves the mean average precision (mAP) across different classes. Furthermore, experiments using masked BEV maps showcase the robustness of the proposed model to degraded sensor inputs. Overall, the proposed full-size model achieves a

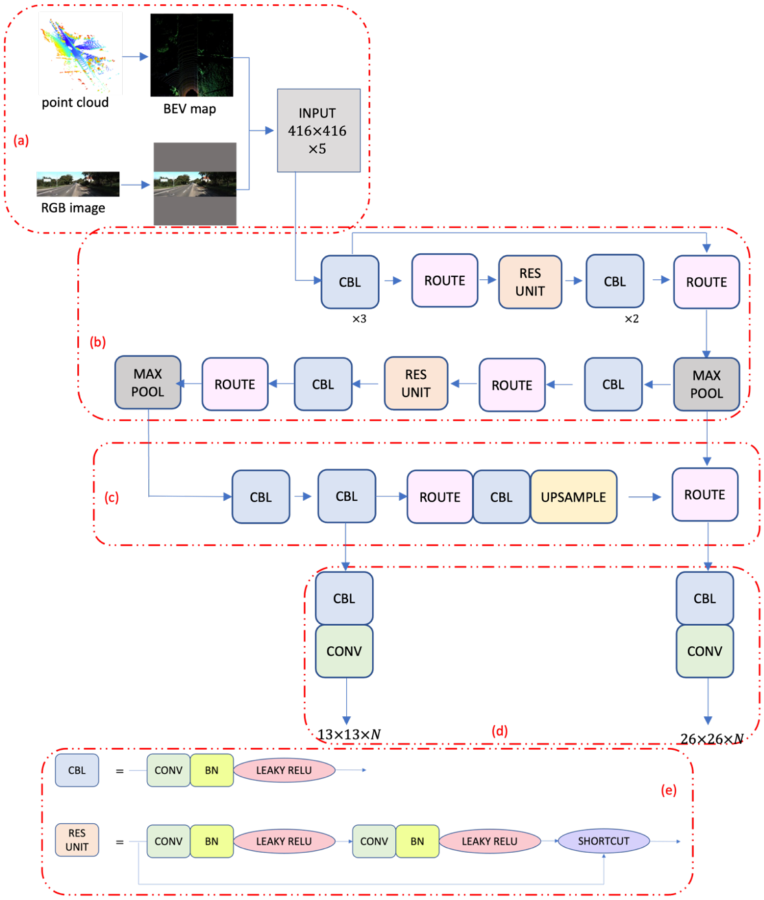


Figure 10: Proposed 3D object mini-detector architecture: (a) preprocessing converts point cloud and rescaled RGB image into a 5-channel 2D BEV map; (b) reduced backbone of proposed mini detector (DarkNet-19); (c) mini FPN; and (d) detection head with output prediction at 2 different scales; with (e) details of the respective structure of CBL (top) and Res Unit (bottom)

Table 4: Detection framerate, precision, recall, AP and F1 estimated on each class and overall mAP for the proposed mini detector compared with the full-size detector

	Class	Mini detector	Full-size detector
Framerate (FPS)		**158.97**	**46.4**
Precision	Car	89.22	90.65
	Pedestrian	47.83	63.89
	Cyclist	68.74	79.51
Recall	Car	95.72	98.68
	Pedestrian	62.31	93.17
	Cyclist	87.72	95.24
AP	Car	93.71	97.94
	Pedestrian	69.08	82.72
	Cyclist	85.44	90.13
F1	Car	92.47	94.50
	Pedestrian	38.38	75.80
	Cyclist	78.32	86.67
mAP (%)		**82.74**	**90.26**

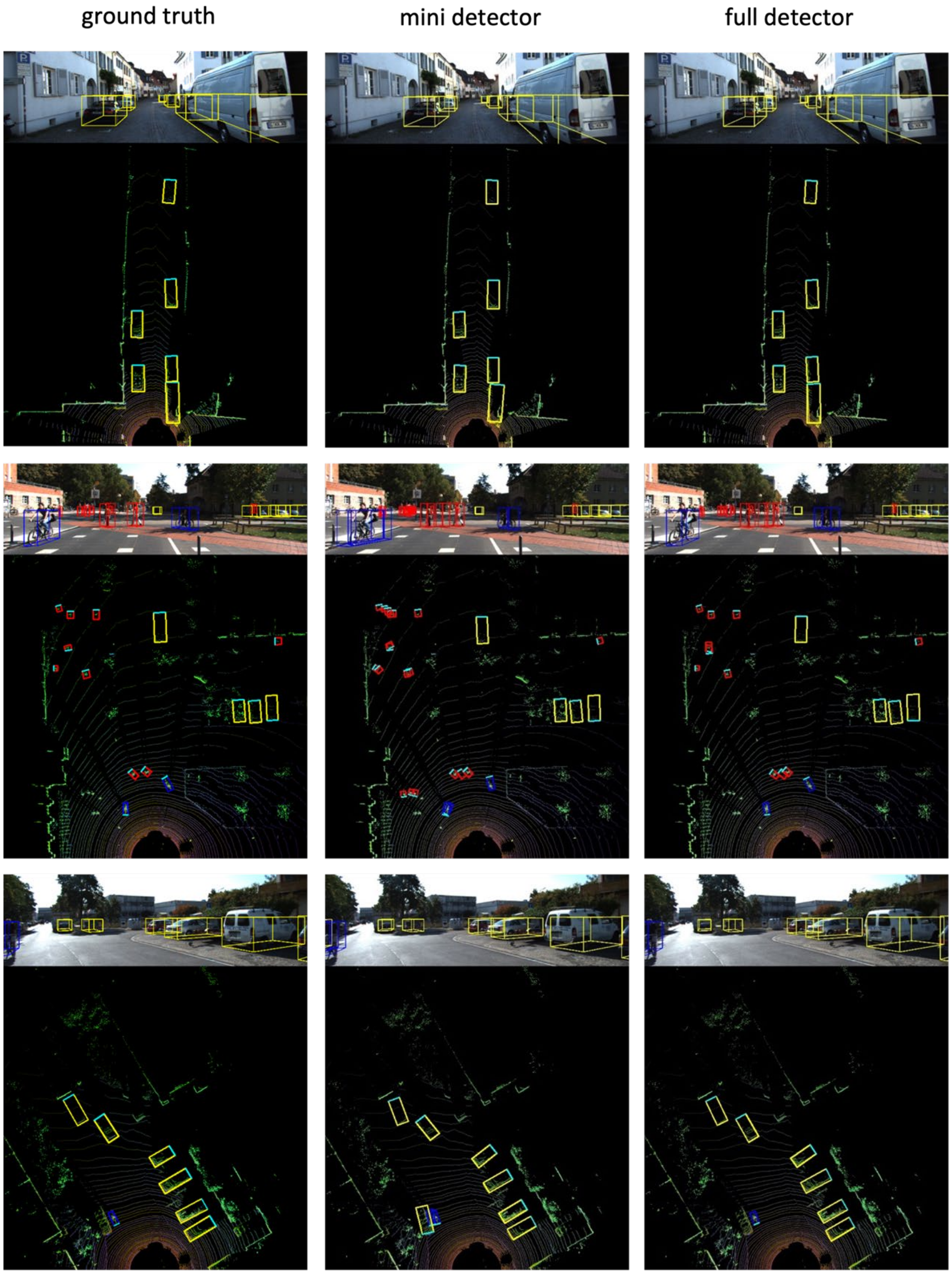


Figure 11: Three sample test cases comparing the ground truth (left), with results of the mini detector (center), and results of the full-size detector (right), with B-Boxes [yellow = car; red = pedestrian; blue = cyclist] superimposed over front forward RGB image (upper part) and over corresponding BEV map (lower part)

Table 5: Comparison of performance between lightweight detection models

Model	Size	Nb of trained parameters	Framerate (FPS)	mAP (%)
Tiny YOLO [16]	60.5 MB	-	133	57.1
Tiny SSD [21]	2.3 MB	1.13 M	-	61.3
Proposed mini detector	*44.9 MB*	*7.19 M*	***158.97***	***82.7***

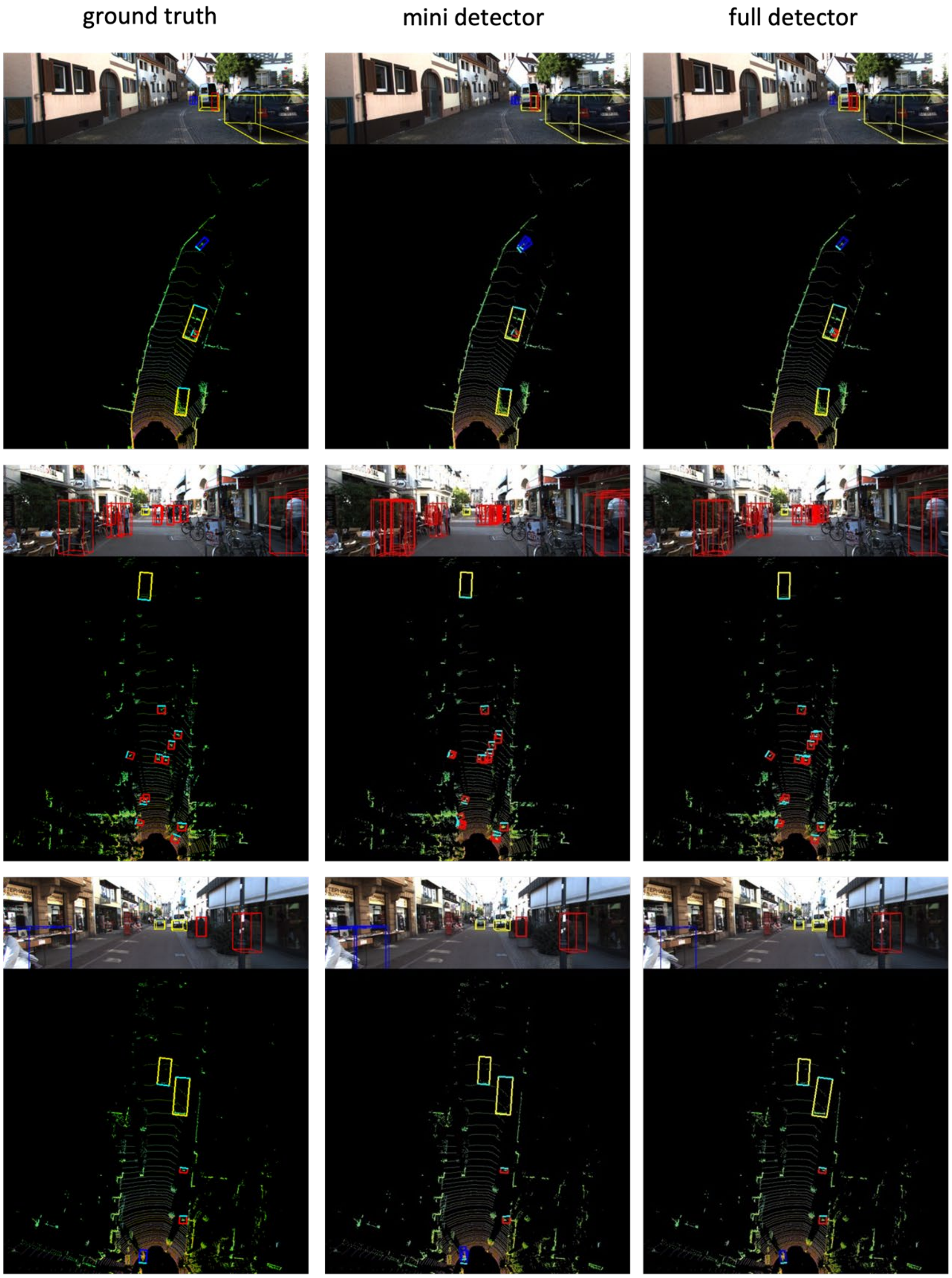


Figure 12: Additional sample cases comparing the ground truth (left), with results of the mini detector (center), and results of the full-size detector (right), with B-Boxes [yellow = car; red = pedestrian; blue = cyclist] superimposed over front forward RGB image (upper part) and over corresponding BEV map (lower part)

detection framerate of up to 46.4 frames per second (FPS) in a single GPU-based implementation with mAP exceeding 90%. Experimental results indicate that the model adapts well to real-life autonomous driving scenarios with varying levels of occlusions.

To explore faster and lighter detection models suitable for real-time and embedded vehicle applications, a mini detector is also introduced. By integrating a lightweight deep learning detector into the 3D data processing domain and leveraging key concepts from the full-size detector, a compact 19-layer network model is developed for 3D object detection and recognition, achieving mAP above 82%. Experiments demonstrate that compared to the full-size detector, the mini detector requires approximately 2/3 of the training time and 1/3 of the testing time. This trade-off between processing time and accuracy allows for effective performance in time-critical applications. The proposed mini detector also shows superior performance in terms of both detection accuracy and framerate compared to other lightweight 3D detectors.

While this research brings significant contributions to 3D object recognition, some limitations remain and open areas for future research. First, it will be beneficial to develop a self-optimizing training model that can automatically adjust training parameters and feature maps to improve the generalization ability of the detectors. This will allow to adapt better to different operational conditions, such as occlusion, low resolution, and varying scene complexity. Second, a sensor-independent fusion framework is essential to ensure the safety of autonomous vehicles. Further research is needed to explore the signal coupling issues that may arise when fusing LiDAR scanner and camera inputs, especially in safety-critical environments.

Moreover, addressing the imbalance in the training dataset is crucial for improving object recognition accuracy. While we employed focal loss to reduce bias, there is still room for improvement, particularly in detecting cyclists and pedestrians, which are as important as detecting cars in autonomous driving scenarios. Future research will investigate methods to make the model less sensitive to the number of training samples or adjust the training dataset to improve balance in the number of samples from different classes.

References

[1] J. Beltrán, C. Guindel, F. M. Moreno, D. Cruzado, F. Garcia, and A. De La Escalera, "Birdnet: A 3D Object Detection Framework from Lidar Information," 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2018.

[2] Y. Chen, S. Liu, X. Shen, and J. Jia, "Fast Point R-CNN," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9775-9784, 2019.

[3] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-View 3D Object Detection Network for Autonomous Driving," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1907-1915, 2017.

[4] A. Geiger, P. Lenz, and R. Urtasun, "Are We Ready for Autonomous Driving? The Kitti Vision Benchmark Suite," *Proceedings of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354-3361, 2012.

[5] C. He, H. Zeng, J. Huang, X.-S. Hua, and L. Zhang, "Structure Aware Single-Stage 3D Object Detection from Point Cloud," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11873-11882, 2020.

[6] L. Kang and P. Payeur, "3D Objects Detection and Recognition from Color and LiDAR Data for Autonomous Driving," *Proceedings of 38th International Conference on Computers and Their Applications,* 91:42-55, 2023.

[7] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D Proposal Generation and Object Detection from View Aggregation," *Proceedings of 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1-8, 2018.

[8] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast Encoders for Object Detection from Point Clouds," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12697-12705, 2019.

[9] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-Task Multi-Sensor Fusion for 3D Object Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7345-7353, 2019.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal Loss for Dense Object Detection," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2980-2988, 2017.

[11] J. Mao, Y. Xue, M. Niu, H. Bai, J. Feng, X. Liang, H. Xu, and C. Xu, "Voxel Transformer for 3D Object Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 3164-3173, 2021.

[12] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum Pointnets for 3D Object Detection from RGB-D Data," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 918-927, 2018.

[13] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-7271, 2017.

[14] J. Redmon and A. Farhadi, "Yolov3: An Incremental Improvement," *arXiv preprint arXiv:1804.02767,* 2018.

[15] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression," *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 658-666, 2019.

[16] K. C. Saranya, A. Thangavelu, A. Chidambaram, S. Arumugam, and S. Govindraj, "Cyclist Detection Using Tiny Yolo v2," *Soft Computing for Problem Solving*, 1057:969-979, 2020.

[17] M. Simon, S. Milz, K. Amende, and H.-M. Gross, "Complex-Yolo: An Euler-Region-Proposal for Real-Time 3D Object Detection on Point Clouds," *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pp. 197-209, 2018.

[18] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-Net: Multimodal Voxelnet for 3D Object Detection," *2019* International Conference on Robotics and Automation (ICRA), 2019.

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All You Need," *Advances in Neural Information Processing Systems,* 30:5998-6008, 2017.

[20] Z. Wang and K. Jia, "Frustum Convnet: Sliding Frustums

to Aggregate Local Point-Wise Features for Amodal 3D Object Detection," 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 1742-1749, 2019.

[21] A. Womg, M. J. Shafiee, F. Li, and B. Chwyl, "Tiny SSD: A Tiny Single-Shot Detection Deep Convolutional Neural Network for Real-Time Embedded Object Detection," 15th Conference on Computer and Robot Vision (CRV*),* pp. 95-101, 2018.

[22] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely Embedded Convolutional Detection," *Sensors,* 18:3377, 2018.

[23] B. Yang, W. Luo, and R. Urtasun, "Pixor: 2017 Real-Time 3D Object Detection from Point Clouds," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7652-7660, 2018.

[24] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-Dense 3D Object Detector for Point Cloud," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1951-1960, 2019.

[25] Z. Zhang and M. Sabuncu, "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," Advances in Neural Information Processing System*s*, 2018.

[26] Y. Zhou and O. Tuzel, "Voxelnet: End-to-End Learning for Point Cloud Based 3D Object Detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4490-4499, 2018.

Lian Kang (photo not available) is a software engineer specializing in software development and artificial intelligence. She holds a Master's degree in Electrical and Computer Engineering from the University of Ottawa, with a focus on machine learning algorithms for computer vision tasks. Her research published in CATA 2023 and IJCA has significantly contributed to advancements in the field of object detection. With a strong background in software engineering and expertise in machine learning, Lian is passionate about leveraging AI technologies to solve real-world problems and drive innovation. Her dedication to cutting-edge research and practical applications makes her a valuable contributor to the field of artificial intelligence.

Pierre Payeur (photo not available) received the Ph.D. degree in Electrical Engineering from Université Laval, Canada. Since 1998, he is a Professor at the School of Electrical Engineering and Computer Science, University of Ottawa. He is the Director of the Sensing and Machine Vision for Automation and Robotic Intelligence research laboratory and a co-founder of the Vision, Imaging, Video Processing and Autonomous Systems research laboratory. He also served in various academic leadership roles while being extensively involved in industrial and international collaborations. His research interests include machine vision, 3D modeling, tactile sensing, automation, manipulator and mobile robotics, and computational intelligence for man-machine interfaces.

Fake News Detection System using BERT and Boosting Algorithm

Raquiba Sultana* and Tetsuro Nishino*
The University of Electro-Communications, Tokyo, JAPAN

Abstract

False information can proliferate and cause significant issues on social media platforms. To minimize the harm caused by false information, understanding its sensitivity and content is essential. This research analyzes the characteristics of human expression and, based on the results, successfully detects fake news by implementing different machine learning models. To identify false information on the Internet, we propose an ensemble model based on transformers. First, various text classification tasks were conducted to understand the contents of false and true news about COVID-19. The proposed hybrid ensemble learning model utilizes these results. The results of our analysis were encouraging, demonstrating that the proposed system can identify false information on social media platforms. All the classification tasks were validated and displayed outstanding results. The final model exhibited excellent accuracy (0.99) and f1 score (0.99). The Receiver Operating Characteristics (ROC) curve showed that the true-positive rate of the data in this model was close to one, and the Area Under the Curve (AUC) score was very high at 0.99. Thus, it was demonstrated that the proposed model effectively identified false information online.

Key Words: NLP, deep learning, text classification, BERT, boosting algorithm.

1 Introduction

The use of social media has steadily increased in recent years. Most Internet users are frequently active on websites such as Facebook, Instagram, and Twitter. Social media users were forecast to number 3.6 billion in 2020; by 2025, that number is projected to rise to 4.41 billion [8]. People frequently rely on social media for daily news. As a result, social media has become the center for spreading false information. The proliferation of fake news has become a global issue, especially during the COVID-19 pandemic. Due to fear of COVID-19, people are more likely to believe false information.

News that is false and disseminated through social media or news outlets is called fake news. In mass media, information accuracy is occasionally compromised to boost revenue. As a result, readers might be misled, and false information might be disseminated regarding politics, religious affiliations, branding, and financial services [35]. False information is propagated to attract public attention, making people more vulnerable to security attacks and harmful social and political issues. This may explain why the current era is defined as the "post-truth" era [24].

Daily news consumption alters how we see the world. The proliferation of false news has jeopardized the integrity of journalism and media. Governments and businesses have traditionally taken measures to define, recognize, and halt the spread of fake news as key goals. Nevertheless, millions receive falsified information daily, which is pervasive on social media. By fostering prejudice and intolerance, misinformation prepares the path for enduring issues.

Many aspects of society have suffered significant damage owing to fake news. For instance, in the stock market, a false story about the parent company of United Airlines declaring bankruptcy in 2008 resulted in a decrease in the stock price by 76% in a matter of minutes, a closing price that was 11% below the previous day, and the negative effect lasted for more than six days [7].

The concept of fake news came into the limelight during the 2016 United States presidential election, and the subsequent social, political, and economic damage caused by the online transmission of misinformation has been well discussed. The prevalence of social media, where spreading false information can easily be done, has worsened this issue. This is frequently carried out to deceive those who believe the news and accomplish economic and political milestones. In addition, the mainstream media has become increasingly biased, and yellow journalism has become more common. Elections, democracy, war, and conflict are the main topics of political news.

In traditional media, politically biased reporting and pulling a predetermined line are frequently used to win over the public. Although such reporting does not spread factually incorrect information, it frequently presents incomplete information to deceive the public and further complicit political interests. Many misleading and inappropriate claims concerning the SARS-CoV-2 novel coronavirus (COVID-19) have been made in conjunction with the virus's outbreak, notably on social media [19]. The World Health Organization (WHO) warned about an ongoing "infodemic", or an excess of information, especially false information, during the pandemic due to the propagation of false information [16].

Since the outbreak of COVID-19, there have been numerous claims that the illness may be cured, including that consuming methanol, ethanol, and bleach can protect one against COVID-19 [38]. The WHO (World Health Organization) had to issue a warning to people not to consume these poisonous substances as a consequence [2]. Political leaders such as President Donald Trump endorsed this assertion, sparking controversy. He frequently described this disease as the Wuhan or China

[*] Graduate School of Informatics and Engineering, Email: sultana.ra@uec.ac.jp, nishino@uec.ac.jp

virus. In response, Asians had been targeted for their race in America. The spread of racial hate crimes is a direct result of misleading information.

Another well-linked hoax involves the 5G network. A rumor that the network was spreading the coronavirus or disrupting human immunity systems first appeared at the start of the lockdown. There were concerns that people ignited communication masts on fire across the UK as a result of the false reports. According to a spokesperson for the Mobile UK industry group, "more than 50" of these arson attacks occurred [4]. Rumors regarding the coronavirus vaccine also spread globally. Numerous studies have examined the relationship between coronavirus vaccine hesitancy and fake news. Anti-vaccine groups tried to demotivate mass people with their far-fetched conspiracy theories [17]. One well-known conspiracy theory claims vaccines permanently damage DNA or alter genes [1]. This myth was only about messenger RNA (messenger-RNA) vaccines, as they implement a genetic approach. These are examples of the spread of fake rumors in recent years. It is increasing alarmingly and requires immediate action to prevent the spread of fake information online.

Fake news creators frequently combine facts from reliable news sources with false materials to purposefully or inadvertently mislead readers. It is increasingly viewed as dangerous to democracy, public peace, and free speech and can confuse people and spark unrest. Many websites have taken on the responsibility of debunking and dismissing rumors and claims, especially those that receive thousands of views and likes before being proven false. A near real-time response is essential to prevent fake information from spreading among online users. Fact-checking websites frequently cannot verify the accuracy of all the latest information fast enough. Identifying fake news aims to save time and effort when examining news veracity[33].

The US 2016 election was questionable for many people as it caused a lot of fake news to spread online. Many researchers have attempted to determine fake news patterns during election periods. Recently, we have proposed a broad framework [27] which might be used in future elections worldwide to help people make better decisions in recognizing news deception and identifying an author's hidden bias. To conduct this study, the researchers built a dataset of 200 tweets about "Hilary Clinton" and conducted a truthfulness assessment. They started by "text normalizing" tweets, examined feature extraction techniques to categorize news, conducted a thorough linguistic analysis of tweets, and extracted the bag-of-words to find observable patterns, and then used the k-nearest neighbor algorithm to distinguish between polarized and credible news. They then discussed the outcomes of implementing the KNN algorithm, interconnected research domains, and future research directions for building an ideal model for a fake news detection system around social media before quantifying the success rate of the proposed framework.

The first step in detecting and preventing the spread of disinformation is understanding the information contained in it. For example, writing patterns, emotions, expression styles, and grammatical accuracy must be analyzed. In other words, it is necessary to identify standard patterns throughout the story. This current study aims to analyze the characteristics of fake and real news. Based on these characteristics, we determine the similarities and differences between the two news types. Recently, several studies have been conducted on this topic. For example, a Naive Bayes classifier was proposed and implemented for spam filtering via emails [13]. This research used Buzzfeed dataset and collected data from three major Facebook pages and three political news pages (Politico, CNN, and ABC News). The model exhibits a classification accuracy of 75.40%.

In another study, we proposed a hybrid fake news detection system focusing on BERT and Ensemble Learning models [28]. This study aimed to analyze the characteristics of fake news by implementing text classification tasks and detecting fake news using an ensemble learning model. These results were impressive. The accuracy score was 0.97, and the f1-score was 0.98.

2 Related Works

Several deep-learning-based methods that perform well on various datasets have been proposed to diminish the online spread of fake news. A recent study proposed a hybrid CNN model that integrated metadata with the text [34]. The authors sought to demonstrate that a hybrid approach could enhance text-only deep learning models. The results of the hybrid CNN were compared with those of support vector machines (SVM), logistic regression, Bi-LSTM, and CNN. Another study suggested an automatic fake news detection system based on a multi-perspective speaker profile [20].

The authors proposed a novel approach for integrating speaker profiles into an attention-based LSTM model to detect falsified news. The profile information served as an attention factor and additional input data. The system performance was assessed using a dataset from [34], and it was shown that adding speaker profiles significantly enhanced the output. The accuracy of this model on the benchmark dataset was 0.415, which is approximately 14.5% greater than that of the most advanced hybrid CNN model.

Public and academic communities have expressed interest in fake news [25]. Such false information has the potential to affect public perception, giving malicious groups a chance to influence the results of public events such as elections. Because of the high stakes involved, automatically identifying fake news is a vital but complex problem that remains poorly understood. However, three features of false news have been universally acknowledged: the language of the article, user responses it receives, and source users endorsing it. The existing research has mainly concentrated on developing solutions specific to a single attribute, which limits their applicability and success. To mitigate this issue, researchers have proposed a CSI model that consists of three modules: capture, score, and integrate

[10]. A recurrent neural network (RNN) is used in the first module, which is based on responses and text, to record the temporal pattern of user activity in a particular article. To evaluate whether an item is fake, the third module is paired with the second module, which learns the source characteristics based on user behavior. CSI outperforms current models in accuracy tests using real-world data and recovers valuable latent representations of users and articles. In recent years, transformers have become the most widely used deep learning model. It was first introduced in a seminar published by several researchers from Google and the University of Toronto [31].

This is a self-attention-based deep-learning language model. The authors suggested a new, straightforward network architecture based solely on attention mechanisms by rejecting the concepts of recurrence and convolutions. According to experiments on two machine translation tasks, these models exhibited superior quality while being more parallelizable and requiring a significant reduction in training time. Since then, several new transformer models have been proposed. These are the modified versions of the base model. In recent years, transformer models have become extremely popular for fake news detection. Several studies have been published based on this topic.

Another study proposed using a transformer-based ensemble of COVID-Twitter-BERT (CT-BERT) models [12]. The authors described the models utilized, methods used for text preprocessing, and how to add more data. The best-performing model demonstrated a weighted f1 score of 98.69 on the test set. Transformer-based models have been used to perform text classification tasks. BERT, RoBERTa, and CT-BERT have been used successfully. The authors also empirically evaluated the effectiveness of a linear support vector baseline (linear SVC) and various text preprocessing techniques and added additional data. Finally, an ensemble learning technique was used to obtain the average of the above models.

Models built on transformers have successfully identified features of social media news. The TweetEval framework, which evaluates tweet classifications for various tasks, was recently proposed. The benchmark for tweet classification, TweetEval,consists of seven Fundamental Heterogeneous Tasks in Social Media NLP Research. The authors compared various pretraining strategies for language modeling and proposed a strong set of baselines as the starting point.The effectiveness of starting with pretrained generic language models and continuing their training on Twitter corpora was demonstrated in these experimental results [3].

In another study, news articles were analyzed to determine whether they were accurate, partially true, false, or something else [30]. The dataset comprised news articles, titles, and article ratings. The data were preprocessed using TF-IDF vectorization, and several machine-learning techniques were employed to select the most effective classification models. The Gradient Boosting technique outperformed all other models. With the best classification accuracy of 0.57 and the highest f1-macro score (0.54 on the provided dataset, the techniques were interpretive. Other classification models, such as Passive Aggressive Classifiers, Logistic Regression Classifiers, and Random Forest Classifiers, have shown different findings.

Another study demonstrated a straightforward method for detecting false information using a Naïve Bayes classifier [13]. The strategy was implemented as a software system and evaluated using data from Facebook news posts. Given the relative simplicity of the model, the classification accuracy of the test set was approximately 74%, which is reasonable. Several methods, which are also explained in this article, can be used to improve the outcomes. According to the results, artificial intelligence techniques can be used to address the challenge of detecting fake news.

Another study examined the rapid expansion of online news content and established whether the news was true or false [11]. Therefore, this research suggests a mechanism to identify rumors and claims that need to be fact-checked, particularly those that receive thousands of views and likes, before being refuted and debunked by reliable sources. Several machine-learning algorithms have been used to identify and categorize fake news. However, the accuracies of these methods are limited. This current study uses a random forest classifier to distinguish between fake and real news. The selected News Dataset was used to extract twenty-three (23) textual features. Out of twenty-three features, 14 were chosen as the best using four techniques, including chi2, univariate, information gain, and feature importance. The proposed model and other benchmark techniques were assessed using the benchmark dataset with the best features. According to the experimental results, the proposed model performed better in terms of classification accuracy than other machine learning methods like GBM (Gradient Boosting Machine), XGBoost (Extreme Gradient Boosting), and the Ada Boost Regression Model.

Social media and news media spread false information to increase the number of viewers or as part of the psychological competition. To mitigate this issue, another study determines a classification of the ensemble using a set of marked as true and false news articles [15]. This study develops a text-based classification approach using an SVM, Random Forest, and Naïve Bayes, and Decision Tree are used as base learners in Bagging and AdaBoost. The goal is to find an answer that allows the user to classify and filter fake material. Consequently, the authors determine that the best-performing classifiers were AdaBoost–Linear SVM and AdaBoost-Random Forest with an accuracy of 90.70% and 80.17%, respectively.

Fact-checking websites play crucial roles in identifying fake news. The difficult process of identifying fake news aims to save time and effort when examining news veracity. For this reason, another study proposed an approach that could identify possible fake news spreaders on social media as the first step towards preventing fake news from being propagated among online users. Therefore, they conducted different learning experiments from multilingual perspectives: English and Spanish. They evaluated different textual features primarily not tied to a specific language and compared different machine-learning algorithms. The results indicated that language-independent features could be used to distinguish between possible fake news spreaders and users who share credible information, with an average detection accuracy of 78% for English and 87% for the Spanish corpus [33].

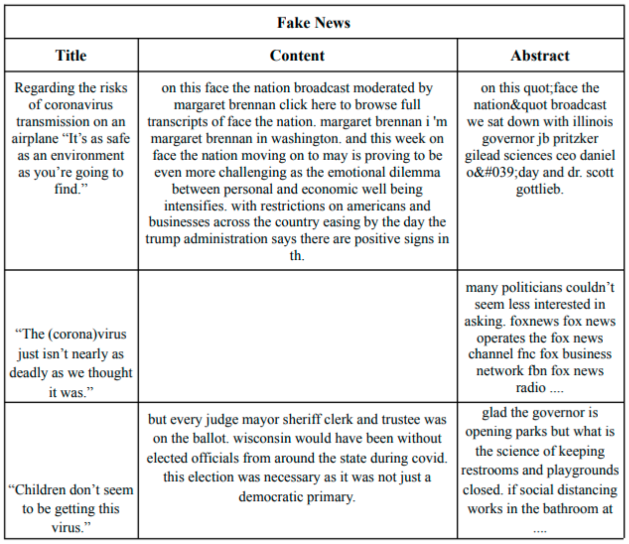
3 Dataset Description

The fake news dataset aims to develop useful features that can distinguish fake news from legitimate news more precisely. Several methods have been developed to acquire news and determine its accuracy. Linguistic traits of news are present in many benchmark datasets for detecting fake information.

The novel coronavirus known as SARS-CoV-2, which was first identified in Wuhan, China, in December 2019, is thought to be the source of COVID-19. SARS-CoV-2 has rapidly spread around the globe. On January 30, 2020, the WHO labeled the outbreak a Public Health Emergency of International Concern [39]. Coughing, shortness of breath, fever, sore throat, and loss of taste or smell are typical COVID-19 symptoms. According to estimates, the incubation period lasts up to 14 days, with a median duration of 5.1 days[18].

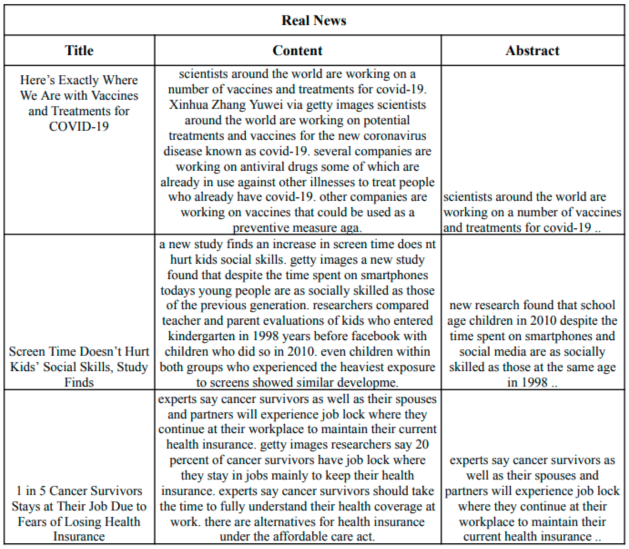
Our society has been affected by COVID-19 for more than two years. The quality of life suffers due to the disruption of supply chains and the impact on the economies of several nations. The disease, infection rates, preventative measures, and vaccinations received daily top-priority news coverage during this time. Because of widespread panic, many people believed that the information shared online was true without checking the source; the spread of false information was almost as bad as the pandemic. This problem is referred to as an "infodemic". Social media sites such as Facebook and Twitter have served as the focal points of this "infodemic". The Co-Aid (COVID-19 Healthcare Misinformation) dataset was chosen for analysis because of this issue [9]. It consists of a variety of healthcare-related COVID-19 data that was obtained from social media.

Information was gathered from December 1, 2019, to September 1, 2020. Three versions were released during this period. In this study, the data were collected from all versions and combined. This information includes news reports, facts, and false information regarding COVID-19. COVID-19, coronavirus, pneumonia, flu9, lockdown, staying at home, quarantine, and ventilators are among the main topics. Most of the posts were gathered from Tiktok, Facebook, Twitter, Instagram, and YouTube. To collect news articles, the author retrieved URLs from several fact-checking websites, including LeadStories, PolitiFact, FactCheck.org, CheckYourFact, AFP Fact Check, and Health Feedback. After obtaining all the URLs of true and fake news related to COVID-19, the authors used newspapers to fetch their corresponding titles, contents, abstracts, and keywords. The original dataset contained 4,251 news articles, 296,000 user interactions, and 926 posts on social media platforms using COVID-19 and ground truth labels. This dataset included information about user engagement on social media as well as information about true and false claims. These were placed in separate files. Only the true and false data were considered in this study. Figure 1 and Figure 2 represent few examples of fake and real news used in the Co-Aid dataset.



Fake News		
Title	**Content**	**Abstract**
Regarding the risks of coronavirus transmission on an airplane "It's as safe as an environment as you're going to find."	on this face the nation broadcast moderated by margaret brennan click here to browse full transcripts of face the nation. margaret brennan i 'm margaret brennan in washington. and this week on face the nation moving on to may is proving to be even more challenging as the emotional dilemma between personal and economic well being intensifies. with restrictions on americans and businesses across the country easing by the day the trump administration says there are positive signs in th.	on this quot;face the nation&quot broadcast we sat down with illinois governor jb pritzker gilead sciences ceo daniel o&#039;day and dr. scott gottlieb.
"The (corona)virus just isn't nearly as deadly as we thought it was."		many politicians couldn't seem less interested in asking. foxnews fox news operates the fox news channel fnc fox business network fbn fox news radio ....
"Children don't seem to be getting this virus."	but every judge mayor sheriff clerk and trustee was on the ballot. wisconsin would have been without elected officials from around the state during covid. this election was necessary as it was not just a democratic primary.	glad the governor is opening parks but what is the science of keeping restrooms and playgrounds closed. if social distancing works in the bathroom at ....

Figure 1: Example of fake news



Real News		
Title	**Content**	**Abstract**
Here's Exactly Where We Are with Vaccines and Treatments for COVID-19	scientists around the world are working on a number of vaccines and treatments for covid-19. Xinhua Zhang Yuwei via getty images scientists around the world are working on potential treatments and vaccines for the new coronavirus disease known as covid-19. several companies are working on antiviral drugs some of which are already in use against other illnesses to treat people who already have covid-19. other companies are working on vaccines that could be used as a preventive measure aga.	scientists around the world are working on a number of vaccines and treatments for covid-19 ..
Screen Time Doesn't Hurt Kids' Social Skills, Study Finds	a new study finds an increase in screen time does nt hurt kids social skills. getty images a new study found that despite the time spent on smartphones todays young people are as socially skilled as those of the previous generation. researchers compared teacher and parent evaluations of kids who entered kindergarten in 1998 years before facebook with children who did so in 2010. even children within both groups who experienced the heaviest exposure to screens showed similar developme.	new research found that school age children in 2010 despite the time spent on smartphones and social media are as socially skilled as those at the same age in 1998 ..
1 in 5 Cancer Survivors Stays at Their Job Due to Fears of Losing Health Insurance	experts say cancer survivors as well as their spouses and partners will experience job lock where they continue at their workplace to maintain their current health insurance. getty images researchers say 20 percent of cancer survivors have job lock where they stay in jobs mainly to keep their health insurance. experts say cancer survivors should take the time to fully understand their health coverage at work. there are alternatives for health insurance under the affordable care act.	experts say cancer survivors as well as their spouses and partners will experience job lock where they continue at their workplace to maintain their current health insurance ..

Figure 2: Example of real news

This information included social media posts and news articles. This study separately combined real and fake news from the entire data collection. A total of 4532 real data and 925 fake data points were utilized in this study. Fake and real data were combined for ease of analysis. Various fact-checking websites validated all the news articles and blog posts. Both true and fake data comprise a statement of the news type (articles/posts, etc.), fact-checking URL, news URL, title, news title, content, abstract, publishing date, and meta-keywords. Information was gathered from news URLs, titles, contents, and abstract columns. The title refers to the news or title of the article and content refers to the content of the news. The abstract refers to a brief description of the article. This research used the title, content, abstract, and URLs among all the information provided. Figure 4 shows a representation of all the analyses performed on the title, content, and abstract. This illustrates the patterns of information dissemination through social media during COVID-19.

4 Methodology

The primary objective of this study was to develop a model that could accurately identify false news on social media. To achieve this, we considered the information gathered from Twitter and examined the characteristics of news articles and social media posts to build a hybrid system for identifying false information on social media. Various text-classification tasks have aided in understanding the characteristics of tweets. This study was influenced by the TweetEval framework [3].

To detect fake news, it is crucial to analyze the data and determine their patterns. This section focuses on the analysis of the data patterns. When training the data for multiple text classification tasks, such as sentiment analysis, emotional analysis, hate speech detection, irony detection, and grammatical analysis, we first investigated the characteristics and patterns of tweets and news articles. All classification tasks were performed using pretrained transformer models from Hugging Face website. Huggingface website offers various pretrained transformer models for different purposes. Available pretrained models can be used for different tasks; such as text classification, image classification, feature extraction, question answering etc. All news items were then rated according to the reliability of their sources. The ensemble learning model was updated based on all the results. The Voting Regressor model received the prediction scores from each classification task as input. The Boosting Ensemble model then received the output score of the voting regressor and the rank score, which predicted whether the news was true or false. Figure 3 depicts the overall process architecture.

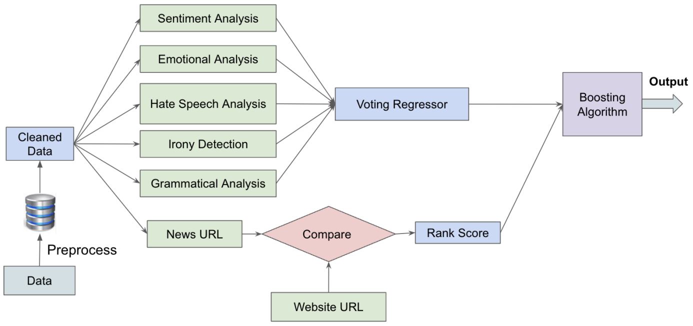


Figure 3: Architecture of proposed model

4.1 Data Pre-processing

The first step was to preprocess the entire dataset. Data preprocessing was the most important step because the raw data were difficult to train. Unprocessed data often yield poor results. This is particularly true when there are large amounts of missing data. The missing values was a crucial issue for the Co-Aid dataset, as many content and abstract data were missing. Those missing from the content columns were handled by inputting the value with the title. However, the missing value in the abstract was replaced with the title. The punctuation was also removed to clean the data. Consequently, the rank scores were normalized using minimum-maximum feature scaling.

4.2 Information Analysis

The data were trained on the basis of all five classification tasks. After training, the prediction scores were transmitted to the ensemble model section. Figure 4 presents the prediction scores of the trained data for all five tasks in the Co-Aid dataset.

The information analysis aimed to determine how people behaved during the COVID-19 pandemic. The "infodemic" era began during the COVID-19 period. They were anxious and believed in anything that could stop the outbreak. Some people made an effort to use this circumstance by spreading false information regarding diseases, prevention, governmental policies, etc. This makes it necessary to examine the patterns of fake news during the pandemic.

- **Sentiment Analysis:** The goal of sentiment analysis is to determine whether tweets are positive, negative, or neutral. The pretrained transformer model CardiffNLP's twitter-roBERTa-base-sentiment-latest [21] was employed to analyze the sentiments [3]. Using this model, the titles, contents, and abstracts were trained. This specific model was pretrained on approximately 124M tweets. The tweets were collected from 2018 to 2021 and fine-tuned for sentiment analysis using TweetEval.

  This pretrained model was applied to the Co-Aid dataset to analyze the sentiments of the data. The findings of the sentiment analysis for COVID-19 are displayed in Figure 4. According to sentiment analysis in Figure 4, neutral news was the most prevalent type, comprising a significant portion of titles, contents, and abstracts. Neutral news accounted for more than 70% of the three cases. However, the prevalence of negative emotions was much lower than that of neutral emotions. Negative emotions ranged from 18% to 24%. Surprisingly, the percentage of positive sentiments is 3%, which is negligible compared to the other cases.
- **Emotion Analysis:** Another text classification task is emotion analysis, which divides data into six categories: anger, fear, joy, love, sadness, and surprise. This assignment aimed to identify the various emotional states in tweets [29]. A pretrained DistilBERT model obtained from the Hugging Face was employed to train the

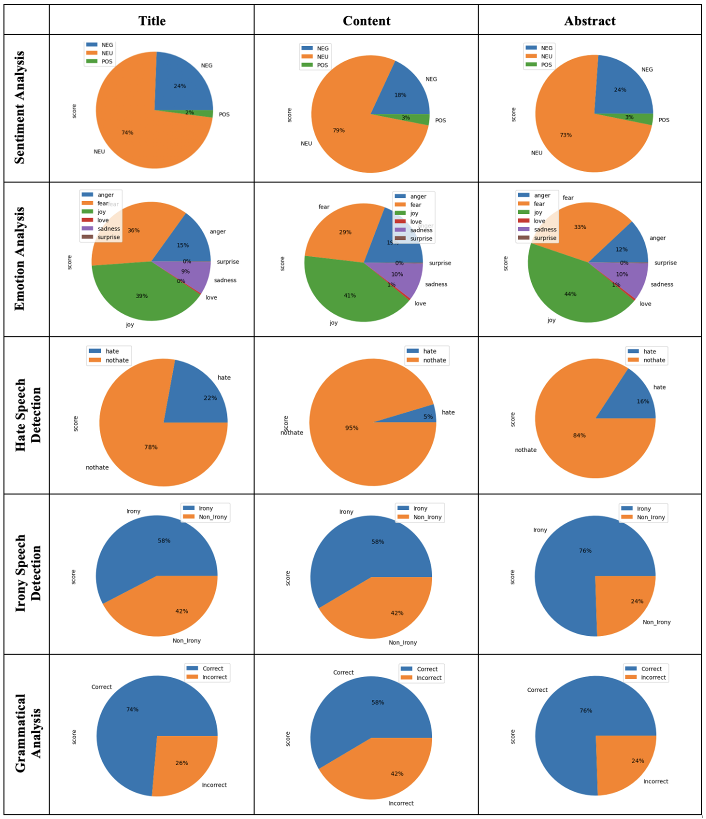


Figure 4: Comparative representation of all the analysis tasks of CoAid dataset

data. ”bhadresh-savani/distilbert-base-uncased-emotion” [26] was employed in this study. Originally, the developer fine-tuned the distilbert-base-uncased model on the emotion dataset [32] using HuggingFace Trainer with specific hyperparameters. The patterns that the posts follow can be explained by emotion analysis.

As illustrated in Figure 4, angry, happy (joyful), and fearful feelings were frequently expressed in news articles. The amount of joyful news was the highest in all three cases. In contrast, the frequency of sad posts was approximately 10%. In contrast, romantic (love) posts were negligible (approximately 1%). The data were gathered at the start of the COVID-19 pandemic, which was characterized by anxiety about the illness and resentment towards the government over measures such as the lockdown. However, when news about vaccines was reported, people felt relieved.

- **Hate Speech Detection:** Hateful content is frequently found in fake news. Although this is true in the case of true news, it is much less likely. Occasionally, people make conscious attempts to spread divisive propaganda. In recent years, bots have been used to spread false propaganda on social media platforms. Therefore, confirming whether the information in news articles is true or false is crucial. HuggingFace’s BERT-based transformer model was used to train the data and spot offensive or hateful content in the news data. During the analysis, we designated offensive information as “hate” and neutral information as “not hate.” The model was pretrained using the HateXplain dataset [22].

  The comparative analysis depicted in Figure 4, shows that most of the cloud data are normal. However, the percentage of abusive or hateful news was too high to be ignored, especially in the title (22%) and abstract (16%).

- **Irony Detection:** Sarcasm is a common way in which people convey emotions. Sarcastic posts contain both accurate and inaccurate information. This ambiguity aids the online dissemination of fake content. Ironic language on social media must be examined to prevent this. This study used the RoBERTa-based transformer model to examine the ironic content in social media. The data were divided into “ironic” and “non-ironic” categories. The results are shown in Figure 4. Surprisingly, most of the posts contained ironic data. The title and content both consist of 58% of the ironic data. This amount was the highest in the abstract (76%). Although there were more ironic posts and news stories, significant percentages of non-ironic posts regarding titles (42%), content (42%), and abstracts (24%) remain.

- **Grammatical Analysis:** The number of people using social media and internet users is proliferating. The number of online newspapers has increased concurrently. Instead of traditional newspapers, people rely on online news portals and social media for news. However, the content quality of online news portals is not sufficiently standardized. These tabloids occasionally circulate false information to boost their audiences. They frequently lack an appropriate editorial board and speak grammatically incorrectly. Therefore, it is important to consider the grammar of any news article.

  To achieve this, a BERT-based model was used to train the data. The Corpus of Linguistic Acceptability (CoLA), which concentrates on the linguistic aspects of texts, was used to pretrain the model. The labels 0 (grammatically incorrect) and 1 (grammatically acceptable) were used to categorize the data [40]. Surprisingly, Fig-4 shows that, aside from the title, most news content and abstracts on social media were grammatically correct. This is true for both social media posts and news articles. The amount of grammatically acceptable data was very high for the title (74%), content (58%), and abstract (76%). This is alarming because newspapers are considered excellent resources for young people learning foreign languages in numerous nations.

After training the data using the aforementioned BERT models, postprocessing tasks were performed on both datasets. The first step was to determine the performance of the models. Therefore, it is crucial to validate all the aforementioned models. As part of the evaluation process, accuracy, precision, recall, and f1 scores were calculated. The final prediction

scores of these models consisted of a label and a score; for example, sentiment analysis yielded positive/negative/neutral labels and their corresponding scores. These two data sets were subsequently combined to yield a final score:
Final Score = Prediction Score + Label Score

On a scale of 0 to 1, the label score represents the frequency of the label among all the data. For example, in the sentiment analysis title of the Co-Aid dataset, negative data comprised 24% of the total data with a label score of 0.24. In contrast, positive data comprised 2% of the total data, giving them a label score of 0.02, and neutral data comprised 74% of the total data, giving them a label score of 0.74. According to the aforementioned formula, if the neutral news had a prediction score of 0.75, the final scores would be 0.68, 0.31, and 0.75. Similarly, if a piece of positive news had a prediction score of 0.5 and a label score of 0.02, its final score would be 0.5 + 0.02 = 0.52. All five participants performed the task. In addition to calculating the final score, it is crucial to validate all classification tasks. All these tasks were validated to verify whether these models functioned per our expectations.

4.3 Rank Score

News websites can be biased or poorly ranked. The rankings of various news websites served as the foundation for the ranking scores. The credibility of a website affects the news quality. For instance, traditional newspapers such as the New York Times rank higher than satirical news websites such as The Onion. Researchers from Stony Brook University developed the Media Rank website to Rank [36]. Six different rankings were employed by the authors:

1. Reputation Rank
2. Popularity Rank
3. Breadth Rank
4. Ads Indicator
5. Spammer Indicator
6. Political Bias

Because the ranking process was incomplete during the composition of this study, only the breadth rank was considered. The reporting of trustworthy news organizations aims to be politically unbiased. Unlike narrow domains with a few repeating entity occurrences, reliable news sources work hard to cover the full spectrum of important news [36]. Consequently, the depths of insight, scope, relevance, clarity, and reporting accuracy are reflected in the breadth of coverage, which is a key indicator of news quality [23]. Based on the number of distinct entities appearing in news reports, breadth rank quantifies the breadth of coverage. This study determined the rank score for each news source using breadth rank.

$$RankScore = 1/BreadthRank \tag{1}$$

It was not possible to obtain the breadth rank of all the news data considered in this study because it did not cover all news websites. The breadth rank was estimated for cases in which it was not available. In particular, the breadth and rank scores of all government websites were estimated to be 1, as we assumed that government websites provide correct information. The rank score was then used in the ensemble learning model after normalization between 0 and 1.

4.4 Ensemble Learning Model

The second half of the experiment was dedicated to ensemble learning. We aimed to develop a stable model that performs well using a supervised machine learning algorithm. However, under certain circumstances, this requirement can be satisfied by multiple models. To address this problem, an ensemble learning model was used to reduce overfitting and increase the model's generalizability. Ensemble learning involves combining several weakly supervised models to create a stronger and more complete supervised model. The fundamental tenet of ensemble learning is that the other weak classifiers correct errors even if one weak classifier makes an incorrect prediction. Therefore, ensemble-learning models are frequently used to combine various fine-tuned models [37]. Two different types of ensemble models were used in the study.

i) Voting Regressor
ii) Boosting Ensemble

i) Voting Regressor: An ensemble machine-learning model called a voting ensemble (or ”majority voting ensemble”) combines predictions from various other models. This method can be applied to enhance the model performance, ideally producing results superior to those of any individual model used in the ensemble. By combining the results from various models' predictions, a voting ensemble operates. This method can be applied to regression or classification. Calculating the average of the model predictions is necessary for regression [5]. When classifying the data, the predictions of each label were added, and the label with the most votes was predicted. This study used a Voting Ensemble for the regression because the average of all input models must be calculated. The final score is transmitted to the Boosting Ensemble Model. The Boosting model was the last one applied to our data.

ii) Boosting Ensemble: Boosting is another type of Ensemble Model. Developing a series of weak models generally increases the prediction power [6]. Each model compensates for the shortcomings of its predecessors. It employs a gradual learning process, an iterative method that aims to reduce the errors of previous estimators. The entire process is sequential, and to make better predictions, each estimator relies on the one before it [14]. Extreme Gradient Boosting, also known as the XGBoost algorithm, is one of the most widely used boosting techniques. The XGBoost algorithm was used to increase the voting regressor's prediction score and determine the final output of the study. The prediction score obtained from the voting regressor and the rank score served as the model's inputs. This entails the ranking and prediction scores of the title, content, and abstract. The result was a binary score of either zero (false) or one. (true). After the completion of this study, the model was validated to determine how well the suggested model would perform.

The previous version of the proposed model used the aforementioned classification tasks. These tasks were implemented using identical pre-trained hugging-face BERT models. The outcome of these classification tasks was the prediction score. The final scores (obtained from the label and prediction scores) were transmitted to the weighted average ensemble model as the input. In contrast, the rank score was calculated for the given news item. The outputs of the weighted ensemble and rank scores are fed into a Stacking Ensemble classifier. The output of the stacked model successfully distinguishes between true and false news items. Output 0 denotes fake news, and output 1 denotes true news. In our previous system, classification tasks were not validated. However, in the present study, these tasks were validated using the Co-Aid dataset. We implement a voting regressor in the proposed model. In previous studies, we implemented a Weighted Average Ensemble model. Previous research used the Stacking Ensemble model; in this study, we replaced the stacked model with the XGboost model.

4.5 Results

The project was implemented using Python version 3.9 and the NVIDIA environment. The proposed solution was employed using PyTorch. The data was cleaned in the beginning. Handling missing values is crucial, because the abstract column contains many missing data points. HuggingFace Transformer models were used to analyze the title, content, and abstract columns. The following transformer models, which are available on the Hugging Face website, were used to calculate the prediction scores:

1) Sentiment Analysis: CardiffNLP-twitter-roBERTa-based-sentiment-latest [21]
2) Emotion Analysis: Bhadresh-Savani-distilbert-based-uncased-emotion [26]
3) Hate Speech Detection: Hate-speech-CNERG-bert-base-uncased-hatexplain-rationale-two [22]
4) Irony Detection: CardiffNLP-twitter-roberta-base-irony [3]
5) Grammatical Analysis: textattack-bert-base-uncased-CoLA [40]

The dataset was trained using the transformer models. The maximum length of the input data was set to 512 for all the models. The default tokenizers from the pretrained models were used in this study. The prediction scores collected from the classification tasks were applied in the second part of the proposed model. All classification tasks were validated, and the accuracy, precision, recall, and f1 scores were calculated. For validation purposes, 4500 data points were used for training, and 957 data points were used for testing the entire dataset. There were three epochs=3, and the batch size was eight. Surprisingly, the results are satisfactory.

The prediction and rank scores were normalized using minimum–maximum feature scaling. Subsequently, a voting regressor ensemble model is applied to the title, content, and abstract columns. The continuous prediction scores for each column were generated as outputs. Subsequently, the performance of the classification task was measured. Due to this purpose, accuracy, precision, recall, and f1 scores were calculated. Table 1 clearly explains the performance measurements of all classification tasks applied to the Co-Aid dataset. The table successfully presents the accuracy, precision, recall and f1-score. The scores were impressive almost everywhere. This implies that the models provide perfect prediction in the majority cases for Co-Aid dataset. The model performed well in Co-Aid dataset. In most cases, the accuracy, precision, recall, and f1-score of the text-classification n tasks were approximately d 99%. In some cases. precision and recall achieved 100% scores. However, there were some exceptions in which the scores were much lower. This is the end of text classification. In the next step, the prediction scores were transmitted to the ensemble learning module.

Table 1: Evaluation of text classification models

	Sentiment Analysis			
Co-Aid	***accuracy***	***precision***	***recall***	***f1-score***
Title	0.999	0.999	1.0	0.993
Content	0.997	1.0	0.996	0.998
Abstract	0.996	0.997	0.997	0.997
	Emotion Analysis			
	accuracy	***precision***	***recall***	***f1-score***
Title	0.979	0.992	0.982	0.987
Content	0.994	1.0	0.993	0.997
Abstract	0.987	0.992	0.992	0.992
	Hate Speech Analysis			
	accuracy	***precision***	***recall***	***f1-score***
Title	0.994	0.997	0.995	0.996
Content	0.994	0.999	0.994	0.996
Abstract	0.817	0.817	1.0	0.89
	Irony Speech Analysis			
	accuracy	***precision***	***recall***	***f1-score***
Title	0.969	0.997	0.965	0.981
Content	0.994	0.997	0.995	0.996
Abstract	0.993	0.995	0.996	0.996
	Grammatical Speech Analysis			
	accuracy	***precision***	***recall***	***f1-score***
Title	0.991	0.999	0.989	0.994
Content	0.989	0.998	0.987	0.993
Abstract	0.972	0.987	0.978	0.983

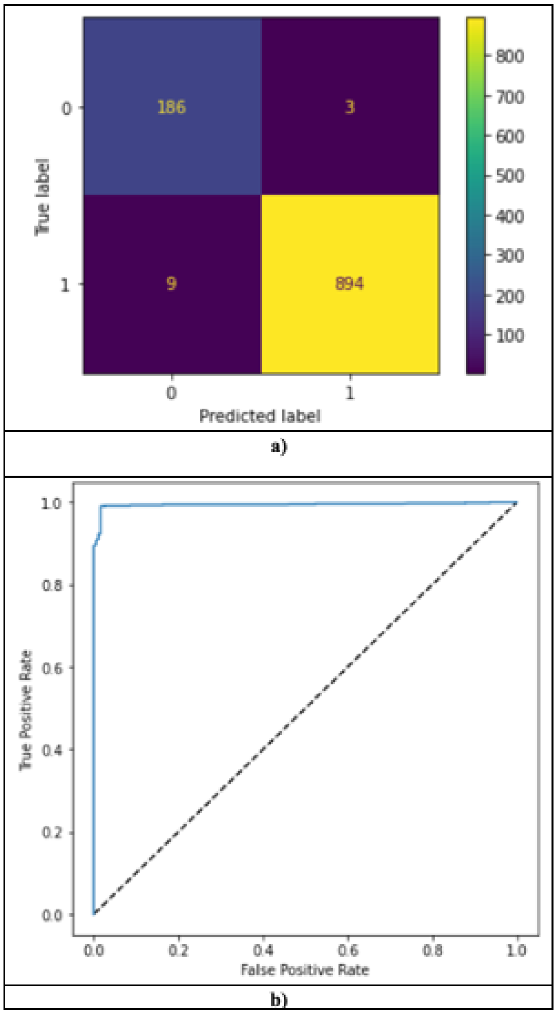
The first step is to apply a Voting Regressor on Title, Content and Abstract. The prediction output needs to be boosted because the results are unsatisfactory. The title, content, abstract, and rank scores were used as inputs for the XGBoost model. The goal of implementing the XGBoost model was to achieve a final score for all news items, including the rank score, and to evaluate the final model. The output column represents the output: Output = 0 if the news is false and output = 1 if it is true. SciKit Learn is employed in the XGBoost model. Approximately 80% of the entire Co-Aid dataset was used for training and 20% for testing. Surprisingly, the Boosting model performed well on the Co-Aid dataset. This successfully boosts the input score, which is the output of the Voting Regressor.

According to figure 5 a), the confusion matrix elaborates more on the prediction employed for the test data of the Co-Aid dataset. Out of the 1092 test samples, our model accurately predicted 893 true and 186 fake data. In contrast, ten true data points were predicted as fake, and three fake data points were predicted as true. This matrix proves that the model accurately predicted most of the time. Consequently, the accuracy, precision, Recall, and f1-Scores were extremely good (accuracy score = 0.98, precision = 1.0, recall = 0.99, f1-score = 0.98, AUC (Area Under the Curve) score = 0.99). By contrast, the ROC AUC curve showed excellent results. Figure 5 b) shows the receiver operating characteristic (ROC) curve of the XGBoost model.

5 Discussion and Conclusion

This study presents an excellent model capable of accurately identifying fake news. However, it only addresses two categories of news: fake and legitimate. Implementing the proposed model on a dataset divided into more than two categories–true, partially true, fake, and partially fake–will be beneficial for obtaining a better understanding. Another issue was the accuracy score and f1-Score of the Hate Speech analysis. In addition to Hate speech analysis, all these models had higher accuracy rates and also f1-score. These issues should be addressed in future studies. Another shortcoming is that this study was implemented only on the Co-Aid dataset. Applying this model to a different dataset can help verify its efficacy. This model can only detect fake online news. We did not consider tracking news propagation or verifying source authenticity. Monitoring the propagation of fake news can help identify the source of the news. This issue will be addressed in future studies.

The comparison between the original model [28] and our suggested models are presented in Table 2. The proposed model performed better than the existing models. The accuracy, precision, recall, f1-score and AUC scores all exhibited improved performance in this new model. The accuracy score was 0.97 in the original model and 0.99 in the XGboost model. f1-score and AUC score is also 0.99 in the proposed model, whereas those were 0.98 in the proposed and original models, respectively. This indicates that the proposed model outperformed the original one. Fake News has become a major issue due to the overwhelming amount of news floating around humans. The spread of fake news has caused enormous harm to society. The proposed model is a small initiative to control false and misleading information. The model is a two-step process in which the initial step is to understand the given information based on different perspectives of human behavior. Prediction scores were successfully calculated by employing pretrained BERT text classification models, such as sentiment analysis, emotion analysis, hate speech detection, irony detection, and grammatical analysis. The model was used to identify fake information in the second step by employing a Voting Regressor, followed by Boosting algorithms. The model performed admirably, displaying high accuracy and an f1 score



a)

b)

Figure 5: a) Confusion matrix and b) ROC Curve of the proposed model for CoAid dataset

Table 2: Comparison between the proposed and original model

Model	***accuracy***	***precision***	***recall***	***f1-Score***	***AUC***
Proposed	0.99	1.0	0.99	0.99	0.99
Original	0.97	0.98	0.98	0.98	0.98

of (0.99) in both cases. The final outcome exhibits the highest AUC rating of 0.99 for the Co-Aid dataset. The TPR rate in this model was close to one, according to the ROC curve, which supports the performance of the proposed model.

Before carefully selecting the final model, several experiments were conducted. The selected combination produced the best outcomes for spotting false information on social media. Calculating the variables for each threshold and plotting them on a plane are required to draw the curve. The performance of the model is illustrated by a curve. Here, the true-positive rate is represented by the blue line, whereas the false-positive rate is represented by the black line. The close proximity of the ROC curve to the axis in the figure indicates the performance of this Boosting model.

References

[1] A. Ahuja, ”COVID-19 Vaccines Myth vs Fact: No Vaccines do not Alter DNA,” https://swachhindia.ndtv.com/covid-19-vaccinesmyth-vs-fact-novaccines-do-not-alter-dna-56612/, NDTV, Feb. 2021.

[2] K. G. Andersen, A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry, ”The Proximal Origin of SARS-CoV-2,” Nat. Med. 26, 450–452. doi: 10.1038/s41591-020-0820-9, 2020.

[3] F. Barbieri, J. Camacho-Collados, L. Neves, and L. Espinosa-Anke, ”Tweeteval: Unified Benchmark and Comparative Evaluation for Tweet Classification,” arXiv preprint arXiv: 2010.12421, 2020.

[4] BBC News, ”Ofcom: Covid-19 5G Theories are “Most Common” Misinformation,” Available, Online: https://www.bbc.co.uk/news/technology-52370616, April 2020, [accessed: Nov. 8, 2022].

[5] J. Brownlee, ”How to Develop Voting Ensembles With Python,” Machine Learning Mastery Website, Online, Available: https://machinelearningmastery.com/voting-ensembles-with-python/, Nov. 2022.

[6] J. Brownlee, ”A Gentle Introduction to the Gradient Boosting Algorithm for Machine Learning,” Machine Learning Mastery Website, Online, Available: https://machinelearningmastery.com/gentle-introduction-gradient-boosting-algorithm-machine-learning/, Aug. 2020 [accessed: Nov. 10 2022].

[7] C. Carvalho, N. Klagge and E. Moench, ”The Persistent Effects of a False News Shock,” *Journal of Empirical Finance* 18(4): 597–615, 2011.

[8] J. Cement, ”Number of Social Media Users 2025,” Statista, Online, Available: https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/, Oct. 2020.

[9] L. Cui, and D. Lee, ”Coaid: Covid-19 Healthcare Misinformation Dataset,” arXiv preprint arXiv: 2006.00885, 2020.

[10] A. Dey, R. Z. Rafi, S. H. Parash, S. K. Arko, and A. Chakrabarty, ”Fake News Pattern Recognition Using Linguistic Analysis,” *In 2018 Joint 7th International Conference on Informatics, Electronics & Vision (ICIEV) and 2018 2nd International Conference on Imaging, Vision & Pattern Recognition (icIVPR)*, IEEE, pp. 305-309, June 2018.

[11] M. Fayaz, A. Khan, M. Bilal and S. U. Khan, ”Machine Learning for Fake News Classification with Optimal Feature Selection,” *Soft Computing*, 26(16): 7763-7771, 2022.

[12] A. Glazkova, M. Glazkov and T. Trifonov, ”g2tmn At Constraint@ aaai2021: Exploiting CT-BERT and Ensembling Learning For COVID-19 Fake News Detection,” *In International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, Cham: Springer International Publishing, pp. 116–127, Feb. 2021.

[13] M. Granik, and V. Mesyura, ”Fake News Detection Using Naive Bayes Classifier,” *In 2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*, IEEE, pp. 900–903, May 2017.

[14] I. Ismiguzel, ”Practical Guide to Ensemble Learning,” TOPBOTS Website, Online, Available: https://www.topbots.com/practical-guide-to-ensemble-learning/, Sep. 2021 [accessed: Nov. 2022].

[15] M. Z. Khan, and O. H. Alhazmi, ”Study and Analysis of Unreliable News Based on Content Acquired Using Ensemble Learning (Prevalence of Fake News on Social Media),” *International Journal of System Assurance Engineering and Management*, 11(2): 145-153, 2020.

[16] N. Kassam, ”Disinformation and Coronavirus,” The Interpreter, Online, Available: https://www.lowyinstitute.org/the-interpreter/disinformation-coronavirus, Lowy Institute, March 2020, [accessed: June 26, 2022].

[17] Y. H. Khan, T. H. Mallhi, N. H. Alotaibi, A. I. Alzarea, A. S. Alanazi, N. Tanveer, and F. K. Hashmi, ”Threat of COVID-19 Vaccine Hesitancy in Pakistan: The Need for Measures to Neutralize Misleading Narratives,” *The American Journal of tropical medicine and hygiene*, 103(2):603, 2020.

[18] S. A. Lauer, K. H. Grantz, Q. Bi, F. K. Jones, Q. Zheng, H. R. Meredith, A. S. Azman, N. G. Reich, and J. Lessler, ”The Incubation Period of Coronavirus Disease 2019 (covid-19) From Publicly Reported Confirmed Cases: Estimation and Application,” *Annals of Internal Medicine*, 172(9):557-582, 2020.

[19] S. van Der Linden, J. Roozenbeek, and J. Compton, ”Inoculating Against Fake News About COVID-19,” *Frontiers in Psychology*, 2928, 2020.

[20] Y. Long, Q. Lu, R. Xiang, M. Li, and C. R. Huang, ”Fake News Detection Through Multi-Perspective Speaker Profiles,” *In Proceedings of the eighth international joint*

conference on natural language processing, Short papers 2:252-256, Nov. 2017.

[21] D. Loureiro, F. Barbieri, L. Neves, L. E. Anke, and J. Camacho-Collados, ”Timelms: Diachronic language models from twitter,” arXiv preprint arXiv:2202.03829, 2022.

[22] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, ”Hatexplain: A Benchmark Dataset for Explainable Hate Speech Detection,” *In Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867-14875, May 2021.

[23] F. Plasser, ”From Hard to Soft News Standards? How Political Journalists in Different Media Systems Evaluate the Shifting Quality of News,” *Harvard International Journal of Press/Politics* 10(2): 47–68, 2005.

[24] A. Qayyum, J. Qadir, M. U. Janjua, and F. Sher, ”Using Blockchain To Rein in The New Post-Truth World and Check The Spread of Fake News,” *IT Professional*, 21(4): 16–24, 2019.

[25] N. Ruchansky, S. Seo, and Y. Liu, ”Csi: A Hybrid Deep Model for Fake News Detection,” *In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797-806, Nov. 2017.

[26] B. Savani, ”Distilbert-Base-Uncased-Emotion,” Online, Available: https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion, Oct. 2022.

[27] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, ”Fakenewsnet: A Data Repository with News Content, Social Context, and Spatiotemporal Information for Studying Fake News on Social Media,” *Big data* 8(3):171-188, 2020, 2020.

[28] R. Sultana and T. Nishino (2022), ”Fake News Detection Using Transformer and Ensemble Learning Models,” *2022 13th International Congress on Advanced Applied Informatics Winter (IIAI-AAI-Winter)*, IEEE, 2022.

[29] L. Tunstall, L. von Werra and T. Wolf, ”Natural Language Processing With Transformers,” O’Reilly Media, Inc., 2022.

[30] R. S. Utsha, M. Keya, Md. Arid Hasan, and M. S. Islam, ”Qword at CheckThat! 2021: An Extreme Gradient Boosting Approach for Multiclass Fake News Detection,” In CLEF (Working Notes), pp. 619-627, 2021.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, and I. Polosukhin, ”Attention is All You Need,” *Advances in Neural Information Processing Systems*, pp. 30, 2017.

[32] A. P. B. Veyseh, F. Dernoncourt, Q. H. Tran, and T. H. Nguyen, ”What does this Acronym Mean? Introducing a New Dataset for Acronym Identification and Disambiguation,” arXiv preprint arXiv:2010.14678, 2020.

[33] I. Vogel and M. Meghana, ”Detecting Fake News Spreaders on Twitter from a Multilingual Perspective,” *In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, pp. 599-606, Oct. 2020.

[34] W. Y. Wang, “Liar, Liar Pants on Fire: A New benchmark Dataset for Fake News Detection,” arXiv preprint arXiv: 1705.00648, 2017.

[35] C. Wardle, and H. Derakhshan, “Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking,” 2017.

[36] J. Ye and S. Skiena, ”MediaRank: Computational Ranking of Online News Sources,” *In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2469–2477, July 2019.

[37] S. Zhou, J. Li, and H. Ding, ”Fake News and Hostile Posts Detection Using an Ensemble Learning Model,” *In Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, Springer International Publishing, Revised Selected Papers 1*, pp. 74-82, Feb. 2021.

[38] World Health Organization, ”Coronavirus Disease (COVID-19) Advice for the Public: Mythbusters,” Online, Available: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters, Aug. 2020, [accessed November 8, 2022].

[39] World Health Organization, ”Statement on the Second Meeting of the International Health Regulations (2005) Emergency Committee Regarding the Outbreak of Novel Coronavirus (2019-ncov),” Online, Available: https://www.who.int/news-room/detail/30-01-2020-statement-on-the-second-meeting-of-the-international-health-regulations-(2005)-emergency-committee-regarding-the-outbreak-of-novel-coronavirus-(2019-ncov), Jan. 2020, [accessed: Oct. 10, 2022].

[40] ”bert-base-uncased-CoLA”, Online, Available: https://huggingface.co/textattack/bert-base-uncased-CoLA, Oct. 2022.

[41] ”Hugging Face – The AI Community Building the Future,”, Online, Available: https://huggingface.co/, [accessed: Aug. 4, 2023].

Raquiba Sultana is a Ph.D. Researcher at The University of Electro-Communications, Tokyo, Japan. She is awarded with Japanese government scholarship, MEXT (Monbukagakusho Scholarship) for pursuing her Ph.D. Her research focuses on Artificial Intelligence and Blockchain. She is enthusiastic about working on applications of artificial intelligence, especially on Natural Language Processing. She also has experience in software engineering in the healthcare industry, and project management in the EdTech field.

Tetsuro Nishino is a Professor at The University of Electro-Communications. He is also a theoretical computer scientist. His research focuses on quantum computations, neural nets, and circuit complexity theory. His primary interest is the computational complexity theory for finding new computation models. He is affiliated with the Information Processing Society of Japan; the Japanese Society for Artificial Intelligence; the Institute of Electronics, Information and Communication Engineers; the Mathematical Society of Japan; the Association for Computing Machinery (ACM).

Secured Communication in Generalized Non DHT-based Pyramid Tree P2P Architecture

Nick Rahimi[*]
University of Southern Mississippi, Hattiesburg, MS

Indranil Roy[†], Ziping Liu[†]
South East Missouri State University, Cape Girardeau, MO

Bidyut Gupta[‡]
Southern Illinois University; Carbondale, IL

Narayan Debnath[§]
Eastern International University; VIETNAM

Abstract

In this paper, we have considered a recently reported 2-layer non-DHT-based structured P2P network. It is an interest-based system. When first reported, peers in each cluster in the architecture used to possess instances of a particular resource type only. It was definitely a very hard restriction practically. Recently, to overcome this restriction a generalized form of the architecture has been reported in which any peer in any cluster can have more than one resource type. This generalized architecture is a little bit more complex than the original one and yet efficiency of each data look-up protocol in it remains the same as in the simpler initial version of the architecture. In this paper, as a continuation of our work in this direction, we have considered security in communication in the generalized architecture. To achieve it, mainly public key-based approach has been used for the different look-up protocols (except for multicasting) because the required number of public-private key pairs is small. However, for multicasting among the cluster-heads, we have followed a hybrid approach, because number of symmetric keys required is just one independent of the size of the multicast group and only the public-private key pair of the root cluster-head is needed.

Keywords: Structured P2P network, residue class, interest-based, non-DHT, secured communication.

1 Introduction

Recent trend in designing structured P2P architectures is the use of distributed hash tables (DHTs) [19, 21, 29]. Such overlay architectures can offer efficient, flexible, and robust service [4, 19, 21, 29, 31]. However, maintaining DHTs is a complex task and needs substantial amount of effort to handle the problem of churn. So, the major challenge facing such architectures is how to reduce this amount of effort while still providing an efficient data query service. In this direction, there exist several important works, which have considered designing DHT-based hybrid systems [13, 17, 28, 32]; these works attempt to include the advantages of both structured and unstructured architectures. However, these works have their own pros and cons. Another design approach has attracted much attention; it is non-DHT based structured approach [4, 8, 18, 22, 25]. It offers advantages of DHT-based systems, while it attempts to reduce the complexity involved in churn handling. Authors in [22] have considered one such approach and have used an already existing architecture, known as Pyramid tree architecture originally applied to the research area of 'VLSI design for testability' [7, 20]. It is an interest-based peer-to-peer system [1, 5, 19-12, 18, 22, 25, 27, 30] with peers of common interest clustered together. Its main focus is to improve the efficiency of data lookup protocols in that a query for an instance of a particular resource type is always directed to the cluster of peers which possess different instances of this resource type. So, success or failure to get an answer for the query involves a search in that cluster only instead of searching the whole overlay network as in the case of unstructured networks. However, that a peer can have only one resource type is a hard restriction practically. To overcome this problem, recently, a generalized form of the architecture [9] has been reported in which any peer in any cluster can possess more than one resource type.

1.1 Our Contribution

In the present work, as a continuation of our research in Pyramid tree p2p network area, we have considered security in communication in the generalized architecture and for this we have preferred public key-based cryptographic approach to redesign/edit our already reported intra- and inter-cluster data

[*] School of Computing Sciences & Computer Engineering.

[†] Department of Computer Science.

[‡] School of Computing.

[§] School of Computing and Information Technology.

look-up protocols, and the broadcast protocol without compromising with the efficiencies of the earlier reported protocols. However, a combination of symmetric and public key-based approach has been considered in designing a very efficient multicast protocol.

The organization of the paper is as follows. In Section 2, we talk briefly about some related preliminaries. For a better understanding of the architecture, refer to our recent publications in this direction. In Section 3, the newly designed secured intra- and inter-cluster data look-up protocols have been presented. In Section 4 we have presented the secured broadcast protocol and Section 5 contains the multicast protocol. Section 6 draws the conclusion.

2 Preliminaries

In this section, we present some relevant findings from our recent works on the Pyramid tree based P2P architecture [15, 22-24] for interest-based peer-to-peer system. Residue Class based on modular arithmetic has been used to realize the overlay topology.

Definition 1. *We define a resource as a tuple <R_i, V>, where R_i denotes the type of a resource and V is the value of the resource.*

Note that a resource can have many values. For example, let R_i denote the resource type 'songs' and V' denote a particular singer. Thus <R_i, V'> represents songs (some or all) sung by a particular singer V'.

Definition 2. *Let S be the set of all peers in a peer-to-peer system with n distinct resource types (i.e. n distinct common interests). Then S = {C_i}, $0 \leq i \leq n\text{-}1$, where C_i denotes the subset consisting of all peers with the same resource type R_i. In this work, we call this subset C_i as cluster i. Also, for each cluster C_i, we assume that C_i^h is the first peer among the peers in C_i to join the system. We call C_i^h as the cluster-head of cluster C_i.*

The overlay network considered is a 2-layer non DHT based architecture [22]. At layer-1, there exists a tree like structure, known as a pyramid tree. It is not a conventional tree. A node *i* in this tree represents the cluster-head of a cluster of peers which possess instances of a particular resource type R_i (i.e., peers with a common interest). The cluster-head is the first among these peers to join the system. Layer 2 consists of the different clusters corresponding to the cluster-heads.

2.1 Characteristics of Pyramid Tree

The following overlay architecture has been proposed in [15, 22-24].

- The tree consists of n nodes. The ith node is the ith cluster head C_i^h. The tree forms the layer-1 and the clusters corresponding to the cluster-heads form the layer-2 of the architecture.
- Root of the tree is at level 1.
- Edges of the tree denote the logical link connections among the n cluster-heads. Note that edges are formed according to the pyramid tree structure [7].
- A cluster-head C_i^h represents the cluster C_i. Each cluster C_i is a completely connected network of peers possessing a common resource type R_i, resulting in the cluster diameter of 1.
- The tree is a complete one if at each level j, there are j number of nodes (i.e,. j number of cluster-heads). It is an incomplete one if only at its leaf level, say k, there are less than k number of nodes.
- Any communication between a peer $p_i \epsilon C_i$ and a peer $p_j \epsilon C_j$ takes place only via the respective cluster-heads C_i^h and C_j^h and with the help of tree traversal wherever applicable.
- Joining of a new cluster always takes place at the leaf level.
- A node that does not reside either on the left branch or on the right branch of the root node is an internal node.
- Degree of an internal non-leaf node is 4.
- Degree of an internal leaf node is 2.

2.2 Residue Class

Modular arithmetic has been used to define the pyramid tree architecture of the P2P system.

Consider the set S_n of nonnegative integers less than n, given as $S_n = \{0, 1, 2,\dots\ (n-1)\}$. This is referred to as the set of residues, or residue classes (mod n). That is, each integer in S_n represents a residue class (RC). These residue classes can be labelled as [0], [1], [2], …, [n – 1], where [r] = {a: a is an integer, a ≡ r (mod n)}.

For example, for n = 3, the classes are:

$$[0] = \{\dots, -6, -3, 0, 3, 6, \dots\}$$

$$[1] = \{\dots, -5, -2, 1, 4, 7, \dots\}$$

$$[2] = \{\dots, -4, -1, 2, 5, 8, \dots\}$$

In the P2P architecture, each integer representing a residue class is the logical (overlay) address of the cluster-head of a cluster. For example, logical address of the first cluster-head is 0, for the second one it is 1, and so on. We use the integers belonging to different classes as the logical (overlay) addresses of the peers with a common interest (i.e., peers in the same cluster) and the number of residue classes is the number of distinct resource types; for the sake of simplicity only the positive integer values are used for addressing. It becomes clear that mathematically any class consists of an infinite number of integers; it means that we do not put any limit on the size of a cluster. In general, number of peers can be too large compared to the number of distinct resource types.

An example of a complete pyramid tree of 5 levels is shown in Figure 1. It means that it has 15 nodes/clusters (clusters 0 to 14, corresponding to 15 distinct resource types owned by the 15 distinct clusters). It also means that residue class with *mod 15* has been used to build the tree. The nodes' respective logical addresses are from 0 to 14 based on their sequence of joining the P2P system.

Each link that connects directly two nodes on a branch of the tree is termed as a *segment.* In Figure 1, a bracketed integer on a segment denotes the difference of the logical addresses of the two nodes on the segment. It is termed as *incremen*t and is denoted as *Inc.,* this increment can be used to get the logical address of a node from its immediate predecessor node along a branch. For example, let X and Y be two such nodes connected via a segment with increment *Inc*, such that node X is the immediate predecessor of node Y along a branch of a tree which is created using *residue class with mod n.* Then, *logical address of Y = (logical address of X + Inc) mod n*.

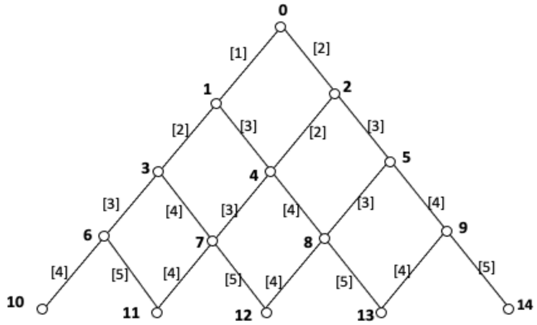


Figure 1: A complete pyramid tree with root 0

Thus, in the example of Figure 1, logical address of the leftmost leaf node = (logical address of its immediate predecessor along the left branch of the root + Inc) mod 15 = (6 + 4) mod 15 = 10.

3 Public Key and Symmetric Key-Based Secured Communication

Before we present the secured protocols the following information from the generalized architecture needs to be recalled. In the generalized architecture each cluster-head maintains a table of information *(TOI)* consisting of tuples corresponding to different cluster-heads. For example, the tuple for cluster-head C_i^h appears as < Res. Code i, IP (C_i^h) >. If a peer p in cluster C_i possesses another distinct resource type, say j, it will appear as a cluster-head as C_j^h in the *TOI.* Same is true if an existing cluster-head C_i^h possesses another distinct resource type, say k, it will appear as a different cluster-head Ck^h However, both Ck^h and Ci^h will have the same IP address but with different overlay addresses k and i respectively. We shall use *TOI* in some of our presented protocols here. Besides, we shall use the broadcast protocol designed for the generalized architecture in our protocols wherever needed. This protocol is known as generalized-broadcast-incomplete protocol. Throughout our presentations, we shall interchangeably use the words 'node' and 'cluster-head'. So, a node on the tree is actually a cluster-head. These are all peers though. However, we strictly use the word 'peer' to represent members of a cluster only to avoid any possible confusion. In addition, we assume that 'resource with type k' and 'resource with code k' mean the same resource.

We have considered public key-based approach for most of the presented protocols in this work. However, symmetric key-based approach can also be used. As an example, we have shown how symmetric key-based approach can be used for intra-cluster data look-up. In addition, we have used a combination of both public and symmetric keys in designing the proposed secured multicast protocol.

In the rest of this section, we shall use the following notations. Public key and private key of the root cluster-head C_0^h are denoted as PU_0 and PR_0 respectively. For any other cluster-head C_i^h, these are PU_i and PR_i respectively. Request for an instance of a resource with code *i* is denoted as *Req-i*; and of course, the IP packet containing *Req-i* will have the requesting peer's identity, that is, its IP address. The corresponding response to *Req-i* is denoted as *Res-i*.

The process of encrypting a message M with a ***key*** is denoted as *E (**key**, M)* and decrypting as *D (**Key**, M).* We have mentioned earlier that resource code is the same as the cluster-head's logical address; for example, if the logical address of a cluster-head C_i^h is i, resource code of the resource owned by the cluster-head as well as by peers in the cluster C_i is also i.

3.1 Public Key Distribution Among Cluster-Heads [16]

We start with a simple method for cluster-heads to know the public keys of all other cluster-heads. Assume that so far there are k number of clusters present in the network, with logical addresses of the existing cluster-heads being 0 to (k-1). That is, the largest resource code currently present in *TOI* is (k-1). In this context note that in the generalized P2P network, it is possible that IP(C_i^h) = IP(C_j^h), i.e., the same peer represents two cluster-heads of two different clusters C_i and C_j containing respectively peers with resource type i and resource type j. Therefore, this peer will have two different logical addresses and two pairs of public key-private key; and the peer will use both the public keys and the private keys of the cluster-heads whenever needed. Same is true for any number of cluster-heads that the peer may represent. When a peer p with instance(s) of some resource type wants to join, at first cluster-head C_0^h checks its *TOI* if the resource type of the joining peer is already present in it. Now one of the following two different situations may arise.

Situation 1: Resource type of peer p does not exist in *TOI*

1. *New peer p contacts C_0^h for joining the network.*
2. *Cluster-head C_0^h assigns the joining peer with the next largest available number for the code; so, the code for p becomes k because TOI contains resource codes from 0 up to (k-1) before peer p joins.*
3. *C_0^h makes entry in the TOI for the new resource code k (which is now the logical address of the newly joining peer p) and the IP address of peer p, because now peer p becomes the cluster-head C_k^h.*
4. *C_0^h and C_k^h exchange their public keys, namely PU_0 and PU_k. Their respective private keys are PR_0 and PR_k. Cluster-head C_0^h updates a list L by including PU_k in it. List L now contains (k+1) different public keys*

corresponding to (k +1) cluster-heads. / List L initially contains only PU_0 of C_0^h / The same peer representing multiple cluster-heads with identical IP addresses but with different resource types, will have unique public key-private key pairs for each such cluster-head

5. *C_0^h performs E (PR_0, L) and executes the generalized-broadcast-incomplete protocol so that each node (cluster-head) on the tree receives a copy of the encrypted list L; each receiving node performs D (PU_0, E (PR_0, L)) to get the recent copy of the list L. / Now each node has the knowledge of the public keys of all existing nodes.*
6. *The above five steps are executed each time a peer with a new resource type contacts C_0^h to join the network.*

Situation 2: Resource type of peer p exists in *TOI*

When a new peer p with some instance(s) of an existing resource type with code, say, *m,* wants to join the system, the following steps are executed to include the peer in the network.

1. *New peer p contacts C_0^h for joining the network*
2. *C_0^h checks with the $(m+1)^{th}$ entry in TOI to get the IP address of C_m^h*
3. *C_0^h sends the IP address of C_m^h to p*
4. *Peer p contacts C_m^h*
5. *C_m^h gives its public key PU_m to peer p / Peer p will use this public key for secured communication inside cluster C_m as well as for inter-cluster data-look-up*

3.2 Symmetric Key Distribution Among Peers in a Cluster

When a new peer with some instance(s) of an existing resource type *m* wants to join the system, the following steps are executed to include the peer in the network.

1. *New peer p contacts C_0^h for joining the network*
2. *C_0^h checks with the $(m+1)^{th}$ entry in TOI to get the IP address of C_m^h*
3. *C_0^h sends the IP address of C_m^h to p*
4. *Peer p contacts C_m^h*
5. *C_m^h and peer p exchange their symmetric keys. / Peer p may use this symmetric key for secured intra-cluster communication in cluster C_m*

Note that if the number of peers in a cluster C_i is N, there will be N distinct symmetric keys, one for each peer; the list containing these corresponding N symmetric keys will be present with the cluster-head C_i^h.

Observation 1. Considering both Situations 1 and 2, total number of required Public-Private key pairs (N') is the number of cluster-heads present in TOI. Some peer may appear as cluster-head multiple times in a generalized situation; thus, depending on the situation this peer may possess more than one public-private key pair.

Observation 2. Number of required symmetric keys (N") for secured communication inside a cluster is the number of peers present in the cluster not counting the cluster-head. This number N" is much larger than N' because number of peers in a cluster is supposed to be very large compared to the number of distinct resource types [6].

Observation 3. Public key-based approach is preferred to symmetric based approach because of much smaller number of distinct keys required in the former.

In all protocols stated in this section, it is assumed that each cluster head has a copy of the list *L* and *TOI.* Besides, it is assumed that peers in a cluster are trustworthy and they are not intruders. An intruder peer is an outsider and does not belong to the P2P network.

3.3 Secured Intra-cluster Data Look-up Protocol

Assume that a peer p' in cluster C_i issues a data look-up request *Req-i* in C_i. Either public key or symmetric key cryptography can be used for secured communication. In the presented protocols, we denote IP address of a peer X as IP(X). We first state secured intra-cluster protocol using public key based cryptographic approach (Figure 2), followed by symmetric key based approach (Figure 3). After that we shall consider secured communication with anonymity.

3.3.1 Use of Public Key Cryptography

```
1. if  $p' = C_i^h$
        $C_i^h$ broadcasts the request Req-i in $C_i$/One hop communication
   .    if  a peer $p^*$ in $C_i$ has the requested response Res-i
            peer $p^*$ unicasts E ($PU_i$, Res-i) to $C_i^h$
            $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-i)) to receive the response Res-i
        else
            Search for Req-i fails
2. else
        p' unicasts Req-i to $C_i^h$
        if  $C_i^h$ has the requested response Res-i
            $C_i^h$ unicasts E ($PR_i$, Res-i) to peer p'
            peer p' performs D ($PU_i$, E ($PR_i$, Res-i)) to receive the response Res-i
        else
            $C_i^h$ broadcasts the request Req-i in $C_i$
3.          if  a peer $p^*$ in $C_i$ has the requested response Res-i
                peer $p^*$ unicasts E ($PU_i$, Res-i) to cluster-head $C_i^h$
                $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-i)) to get Res-i
                $C_i^h$ unicasts E ($PR_i$, Res-i) to the peer p'
                  peer p' performs D ($PU_i$, E ($PR_i$, Res-i)) to receive the response Res-i
4.          else
                Search for Req-i fails
```

Figure 2: Public key based secured intra-cluster data lookup

Remark 1. Only the public-private key pair of the cluster-head is needed to ensure security.

3.3.2 Use of Symmetric Key Cryptography

```
1.  if  p' = $C_i^h$
        $C_i^h$ broadcasts the request Req-i in $C_i$  / One hop communication
    .   if  a peer p* in $C_i$ has the requested response Res-i
            P* unicasts E (SyKey (p*, $C_i^h$), Req-i) to cluster-head $C_i^h$
            / SyKey (p*, $C_i^h$) is the common symmetric key of p* and $C_i^h$
            $C_i^h$ performs D (SyKey (p*, $C_i^h$), E (SyKey (p*, $C_i^h$), Res-i)) to get Req-i
        else
            Search for Req-i fails
2.  else
        p' unicasts its Req-i to cluster-head $C_i^h$
        if  $C_i^h$ has the requested response Res-i
            $C_i^h$ unicasts E (SyKey (p', $C_i^h$), Res-i) to peer p'
                peer p' performs D (SyKey (p', $C_i^h$), E
                (SyKey (p', $C_i^h$), Res-i)) to get Req-i
3.      else
            $C_i^h$ broadcasts the request Req-i in $C_i$
            if  a peer p* in $C_i$ has the requested answer
                peer p* unicasts E (SyKey (p*, $C_i^h$), Res-i) to cluster-head $C_i^h$;
                $C_i^h$ performs D (SyKey (p*, $C_i^h$), E (SyKey (p*, $C_i^h$), Res-i)) to check the response Res-i
                $C_i^h$ performs E (SyKey (p',$C_i^h$), Res-i)
                $C_i^h$ unicasts the encrypted response to peer p'
                peer p' performs D (SyKey (p',$C_i^h$), E ( SyKey (p',$C_i^h$), Res-i)) to receive the response Res-i
            else
4.              search for Res-i fails
```

Figure 3: Symmetric key based secured intra-cluster data lookup

We mentioned earlier that number of required symmetric keys for secured communication inside a cluster is the number of peers present in the cluster not counting the cluster-head. This number may be very large. However, for public key-based approach it is only the public key and the private key of a cluster-head C_i^h that are required for secured intra-cluster communication in that cluster C_i. Because of this very small number of keys required, we have considered public key-based approach for designing secured communication protocols with anonymity.

3.4 Secured Intra-Cluster Communication with Anonymity

Let us first state a general idea to achieve anonymity irrespective of if the protocol used is an intra-cluster or an inter-cluster one. Achieving anonymity in the architecture is a bit tricky because diameter of each cluster is just one. It means that for intra-cluster communication, a requester and a replier are always one hop away from each other. So, logically there is no other peer present between them; this means they cannot hide their respective identities from each other while communicating with each other through request and response. It has led us to present the following simple yet effective solution to achieve anonymity between a requester and a responder inside a cluster. It can be applied to inter-cluster communication as well. The idea is somewhat similar to the idea used in *Mixes*. In general, the idea works as follows.

Let the requesting peer be p' and also let p" be a randomly chosen peer by p' toward the destination. Peer p' sends a request packet directly to peer p". Thus, p" acts as an intermediate forwarding peer of the request packet issued by peer p'. Before unicasting further, peer p" replaces the IP address in the source field in the received packet by its own identity. Thus, p" appears now as if it is the source of the requester to the next peer along a randomly chosen path to the destination.

In the present work, the path of query propagation is termed here as *query-path*. Furthermore, we assume that these peers will remember, wherever applicable, their respective immediate senders' identities and immediate followers' identities along the query propagation path so that a reply can follow *the reverse-query-path* all the way to the requesting peer. Similar approach of replacing addresses is followed along the reverse path while forwarding the corresponding response toward the requesting peer. Therefore, the replier will have no clue about who the original requesting peer is. Similarly, the requester will not know the true identity of the replier.

It may appear that too many address replacements may be needed, but we observe that it is not the case in both intra- and inter-cluster lookup protocols; it is at most two. Recall that diameter of a cluster is just one hop. Therefore, in the present work to achieve anonymity, a request is sent by a requesting peer p_i to its cluster-head C_i^h always via a randomly chosen peer (i.e., using only two hops); hence, source identity is replaced only once in a request packet while it travels to the cluster-head except in the case when the cluster-head itself issues the request; in this later case, only the response packet may go through the replacement process once if at all needed. Similarly, a response packet is always forwarded by the cluster-head C_i^h to the requesting peer p_i using two hops only; thereby replacement occurs only once in the response (reply) packet, irrespective of if the proposed protocol is intra- or inter-cluster protocol. Thus, we observe that only two replacements are sufficient to hide the identities of the requesting peer and the responder from each other. In other words, replacements occur only in one segment (from requesting peer to its cluster-head) of the query-path. Similarly, it occurs only in one segment (from the cluster-head to the requesting peer) of the reverse-query-path. Therefore, the look up latency increases only by two hops.

The public key-based secured intra-cluster data lookup protocol with anonymity appears in Figure 4 below.

```
1. if   $p' = C_i^h$
        $C_i^h$ broadcasts the request Req-i in $C_i$   / One hop communication
        .   / $C_i^h$ always broadcasts a request irrespective of if it is the   actual requester or not; so, no receiving peer has any clue about the true requester
        if   a peer p* in $C_i$ has the requested response
            peer$^*$ unicasts E ($PU_i$, Res-i) to a randomly chosen peer $p^r$ in $C_i$
            peer $p^r$ unicasts Req-i to cluster-head $C_i^h$   /Intruder cannot get Res-i as it does not know $PR_i$
            $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-i)) to get the response Res-i   / $C_i^h$ has no clue about the identity of the true responder
        else
            Search for Req-i fails
2.  else
        p' unicasts Req-i to a randomly chosen peer $p^r$ in $C_i$
        peer $p^r$ unicasts Req-i to cluster-head $C_i^h$   /Req-i is sent along the query-path to $C_i^h$
        / $C_i^h$ does not know the identity of the actual requester
3.      if   $C_i^h$ has the requested response Res-i
            $C_i^h$ unicasts E ($PR_i$, Res-i) to the random peer $p^r$ along the reverse-query-path
            peer $p^r$ unicasts E ($PR_i$, Res-i) to peer p'   / reverse-query-path is followed to reach the true requester
            peer p' performs D ($PU_i$, E ($PR_i$, Res-i)) to receive the response Res-i
            /Anonymity of the responder is preserved
        else
            $C_i^h$ broadcasts the request Req-i in $C_i$
4.          if   a peer p* in $C_i$ has the requested response Res-i
                peer p* unicasts E ($PU_i$, Res-i) to cluster-head $C_i^h$
                /Intruder cannot get Res-i as it does not know $PR_i$
                $C_i^h$ decrypts the information to get Res-i
                $C_i^h$ unicasts E ($PR_i$, Res-i) to the random peer $p^r$ along the reverse-query-path
                peer $p^r$ unicasts E ($PR_i$, Res-i) to peer p'   / reverse-query-path is followed to arrive at the true requester
                peer p' performs D ($PU_i$, E ($PR_i$, Res-i)) to receive the response Res-i
                / p' does not know the identity of the true responder
            else
                Search for Req-i fails
```

Figure 4: Public key based secured intra-cluster data lookup with anonymity

Theorem 1. Anonymity between a requesting peer and a responding peer is preserved.

Proof: In the protocol, if cluster-head is the requester, any response arrives at the cluster-head via a randomly chosen peer in the cluster. Hence, cluster-head does not know the identity of the actual responder. On the other hand, if cluster-head is not the requester, any response is always unicasted finally by the cluster-head to the requesting peer via a randomly chosen peer in the cluster. So, such a requesting peer will not have any clue about the true responder. Similarly, any request from a peer other than the cluster-head always arrives at the cluster-head via a randomly chosen peer. So, the cluster-head has no clue about the true requester. Besides, whenever needed whether the requesting peer itself is the cluster-head or not, the Req-i is always broadcasted in the cluster by the cluster-head. So, no responder will have any clue about the true requester. Hence anonymity between a requesting peer and a responding peer is always preserved. □

The protocol offers secured communication because of two reasons: first, any intruder cannot know the private key of any cluster-head; therefore, when a responding peer unicasts to the corresponding cluster-head its response encrypted with the public key of the cluster-head, no intruder can know the response. Second, even if the intruder comes to learn about the public key of the cluster-head, it will be almost impossible to guess the random peer on the reverse-query-path along which the cluster-head unicasts any response encrypted with its private key: meaning that an intruder cannot identify the path of the reply to the requester making it impossible to look for the reply.

Observation 4. Security with anonymity with symmetric key based approach in intra-cluster data look-up can be achieved using query-path and reverse-query-path in a similar way as in the public key-based approach.

3.5 Secured Inter-Cluster Data Look-up Protocol with Anonymity

We shall illustrate the steps of the protocol using the following inter-cluster communication scenario. Let a peer p* in cluster C_i get an instance of a resource with code m. So, the request *Req-m* should be sent to cluster C_m to get any response related to the query. Note that the cluster-head C_i^h itself may also be the requesting node. Besides, the generalized architecture, the same peer may appear as multiple cluster-heads [26]; in such a case $IP(C_i^h)$ and $IP(C_m^h)$ will be identical even though the two cluster-heads will have different logical addresses in the *TOI*. Therefore, we require to consider all four possible cases involving the two cluster-heads and the requesting peer. The protocol is stated below in Figure 5.

```
if Peer p* $\neq C_i^h$ and IP($C_i^h$) = IP($C_m^h$) / Case 1
        Peer p* unicasts Req-m to cluster-head $C_i^h$ via a query-path in $C_i$   / path is chosen as in intra-cluster protocol
    if   $C_m^h$ has the answer to Req-m
```

```
            $C_i^h$ unicasts E ($PR_i$, Res-m) to the requesting peer p* via the reverse-query-path
            / $C_m^h$ and $C_i^h$ are the same peer with different private keys
            and different logical addresses
            peer p* decrypts with $PU_i$ to get the Res-m
            / anonymity of requester and responder is preserved
        else
            $C_m^h$ broadcasts the request in $C_m$
            if ∃ p' with the answer Res-m
                p' unicasts E ($PU_m$, Res-m) to its cluster-head $C_m^h$
                $C_m^h$ performs D ($PR_m$, E ($PU_m$, Res-m) to get Res-m
                $C_i^h$ unicasts E ($PR_i$, Res-m) to the requesting peer p* via the reverse-query-path in $C_i$
                / $C_m^h$ and $C_i^h$ are the same peer with different private keys
                different logical addresses
                peer p* decrypts with $PU_i$ to get the Res-m   / Anonymity of requester and responder is preserved
            else
                $C_i^h$ unicasts ‘fail’ information to peer p* via the reverse-query-path in $C_i$   / Search for Req-m fails
else
    if Peer p* ≠ $C_i^h$ and IP($C_i^h$) ≠ IP($C_m^h$)   / Case 2
        $C_i^h$ finds from list L the public key $PU_m$ of cluster-head $C_m^h$   / Identifies the cluster with resource type m
        $C_i^h$ unicasts E ($PU_m$, Req-m) to cluster-head $C_m^h$   / Inter-cluster communication
        $C_m^h$ performs D ($PR_m$, E ($PU_m$, Req-m)) to get Req-m
        if $C_m^h$ has the answer to Req-m
            It unicasts E ($PU_i$, Res-m) to cluster-head $C_i^h$   / Inter-cluster communication
            $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-m)) to get Res-m   / $C_i^h$ receives the response
            $C_i^h$ unicasts E ($PR_i$, Res-m) to the requesting peer p* via the reverse-query-path in $C_i$
            peer p* decrypts with $PU_i$ to get the Res-m
            / Anonymity of requester and responder is preserved
        else
            $C_m^h$ broadcasts the request in $C_m$
            if ∃ p'' with the answer Res-m
                p'' unicasts E ($PU_m$, Res-m) to its cluster-head $C_m^h$
                $C_m^h$ performs D ($PR_m$, E ($PU_m$, Res-m) to get Res-m
                $C_m^h$ unicasts E ($PU_i$, Res-m) to cluster-head $C_i^h$   / Inter-cluster communication
                $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-m)) to get Res-m   / $C_i^h$ receives the response
                $C_i^h$ unicasts E ($PR_i$, Res-m) to the requesting peer p* via the reverse-query-path in $C_i$
                peer p* decrypts with $PU_i$ to get the Res-m   / Anonymity of requester and responder is preserved
            else
                $C_m^h$ unicasts ‘fail’ information to $C_i^h$   / Search for Req-m fails
                $C_i^h$ unicasts it to peer p* via the reverse-query-path in $C_i$
    else
        if Peer p* = $C_i^h$ and IP($C_i^h$) ≠ IP($C_m^h$)   / Case 3
            $C_i^h$ finds from list L the public key $PU_m$ of cluster-head $C_m^h$   / Identifies the cluster with resource type m
            $C_i^h$ unicasts E ($PU_m$, Req-m) to cluster-head $C_m^h$   / Inter-cluster communication
            $C_m^h$ performs D ($PR_m$, E ($PU_m$, Req-m)) to get Req-m
            if $C_m^h$ has the answer to Req-m
                It unicasts E ($PU_i$, Res-m) to cluster-head $C_i^h$   / Inter-cluster communication
                $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-m)) to get Res-m   / $C_i^h$ receives the answer to its query
                / $C_i^h$ has no clue about the identity of the original responder;
                that is, $C_i^h$ is not sure if a peer in $C_m$ or $C_m^h$ itself has the
                response
            else
                $C_m^h$ broadcasts the request in $C_m$
                if ∃ p'' with the answer Res-m
                    p'' unicasts E ($PU_m$, Res-m) to its cluster-head $C_m^h$
                    $C_m^h$ performs D ($PR_m$, E ($PU_m$, Res-m) to get Res-m
                    $C_m^h$ unicasts E ($PU_i$, Res-m) to cluster-head $C_i^h$   / Inter-cluster communication
                    $C_i^h$ performs D ($PR_i$, E ($PU_i$, Res-m)) to get Res-m   / $C_i^h$ receives the response
        / $C_i^h$ has no clue about the identity of the original
        responder

        / $C_i^h$ is not sure if a peer in $C_m$ or $C_m^h$ itself has the
        response

                else
                    $C_m^h$ unicasts 'fail' information to $C_i^h$
/ Search for Req-m fails

    else
        if Peer $p^*$ = $C_i^h$ and $IP(C_i^h)$ = $IP(C_m^h)$
/ Case 4

                if $C_m^h$ has the answer to Req-m
                    $C_i^h$ receives the answer Res-m to its query
        / $C_m^h$ and $C_i^h$ are the same peer with different private
        keys

        and different logical addresses

        / $C_i^h$ is not sure if a peer in $C_m$ or $C_m^h$ itself has the
        response
              else
                  $C_m^h$ broadcasts the request in $C_m$
                  if ∃ p' with the answer Res-m
                      p' unicasts E ($PU_m$, Res-m) to its
cluster-head $C_m^h$
                      $C_m^h$ performs D ($PR_m$, E ($PU_m$, Res-m)
        to get Res-m
                      $C_i^h$ receives the answer Res-m to its
        query       / $C_i^h$ has no clue about the identity of the
        original responder;

        / $C_i^h$ is not sure if a peer in $C_m$ or $C_m^h$ itself has the
        response
                  Else
                      Query fails
```

Figure 5: Public key based secured inter-cluster data lookup with anonymity

Theorem 2. In inter-cluster communication, anonymity between a requesting peer and a responding peer is preserved.

Justification similar to the one used in Theorem 1 can be used here to prove that anonymity of both requesting and responding peers is preserved.

Observation 5. Only the public and private keys of two cluster-heads are necessary for secured inter-cluster data look-up with anonymity.

Remark 2. For both intra- and inter-cluster data look-ups with anonymity, the source address in a packet needs to be replaced at most twice considering both propagation of a request and propagation of a response.

Remark 3. Data look up latency increases at most by two hops only.

It may be noted that in none of the above two protocols there is a need for any cluster-head to replace the source address in a received packet by its own identity.

4 Secured Broadcast in Pyramid Tree

Let node X be the broadcast source for a message M. The source may be a peer p in some cluster C_i or itself is the cluster-head C_i^h. In the former case, X will first unicast its broadcast message M to its cluster-head C_i^h. which will then unicast M to the root cluster-head C_0^h. This root will actually execute broadcasting in the tree.

In the latter case, C_i^h will unicast its broadcast message to the root which will then broadcast in the tree. Existence of virtual neighborhood [24] is the reason behind the root broadcasting instead of C_i^h. The *Generalized-Broadcast-Incomplete protocol* has appeared in [IJCA June-23]. In this work, we shall incorporate security in this protocol. The following three possibilities need be considered in designing the secured protocol: (1) X is in C_i, but $X \neq C_i^h$, (2) $X = C_i^h$, and (3) $X = C_0^h$

4.1 Protocol Secured Generalized-Broadcast

```
1.   if  X is in $C_i$, but $X \neq C_i^h$
             X unicasts the message encrypted with public key $PU_i$
of $C_i^h$ to cluster-head $C_i^h$
             $C_i^h$ decrypts the message M with its private key $PR_i$
             $C_i^h$ encrypts the message M with its private key $PR_i$
             $C_i^h$ unicasts the encrypted message to $C_0^h$

             $C_0^h$ decrypts the encrypted message with $C_i^h$'s public
key $PU_i$
                                                               /
authenticates the broadcast source
             $C_0^h$ encrypts the message M with its private key $PR_0$
       else
  2.     if  $X = C_i^h$
               $C_i^h$ encrypts the message M with its private key $PR_i$
               $C_i^h$ unicasts the encrypted message to $C_0^h$
               $C_0^h$ decrypts the encrypted message with $C_i^h$'s
public key $PU_i$
                                                               /
authenticates the broadcast source
               $C_0^h$ encrypts the message M with its private key $PR_0$
       else
  3.     if  $X = C_0^h$
               $C_0^h$ encrypts the message M with its private key $PR_0$
/ $X = C_0^h$
  4.     $C_o^h$ executes Generalized- Broadcast-Incomplete
protocol
                         / A receiving cluster-head $C_m^h$ decrypts
the received encrypted message
                          with the public key $PU_0$ of the root to
get the message M
```

```
5. if   broadcasting is needed in any cluster $C_j$
        $C_j^h$ encrypts the message M with its private key $PR_j$
        $C_j^h$ broadcasts in cluster $C_j$              / one hop communication
        each receiving peer in $C_j$ decrypts the encrypted message with $PU_j$ to get M
```

Figure 6: Secured generalized-broadcast protocol

In this protocol step 5 will not be executed if broadcasting involves only the cluster-heads, for example when the root cluster-head broadcasts a copy of the updated *TOI* to all other cluster-heads. In this case, in the protocol, only the public-private key pair of the source cluster-head (if $X \neq C_0^h$) and that of the root are needed. If broadcasting is done in the clusters as well as in step 3 above, the total number of public-private key pairs is the number of the cluster-heads present in the tree.

5 Secured Multicast to a Group of Cluster-Heads

We will use some idea from Protocol Independent Multicasting – Sparse Mode (PIM-SM) for multicasting [2] in WANs. The reason for this is that in a Pyramid tree P2P network, number of cluster-heads is very small compared to the total number of peers in the network because the total number of distinct resources is very limited [6]. Therefore, any multicast group of cluster-heads will definitely have a smaller size. According to PIM-SM a core is needed and obviously the root cluster-head C_0^h becomes the logical choice as the core irrespective of its membership for a given multicast group G.

We shall use a hybrid approach based on symmetric and public keys as stated below. Multicasting has two phases: in the first phase a group symmetric key SyKey(G) is selected by the core and the core distributes the key to all member cluster-heads of the group using core's private key; in the second phase, multicast takes place using the symmetric key. We denote a cluster-head C_i^h wishing to join a group G, as C_i^h *(G),* also called a cluster member, a multicast source (also a cluster-head) as X, and X may or may not belong to the group; we also denote a multicast message as M.

```
Phase 1: Distribution of group-symmetric key

1. Each cluster member $C_i^h$ (G) unicasts a join request Req to the core $C_0^h$
   / a branch from $C_i^h$ (G) to core (root) is built as in PIM/SM, thereby forming a multicast tree with core as its root
2. Core $C_0^h$ selects the cluster-symmetric key SyKey(G) and encrypts it with its private key $PR_0$
3. Core $C_0^h$ unicasts the encrypted key SyKey(G) to each $C_i^h$ (G),
4. Each $C_i^h$ (G) decrypts with core's public key $PU_0$ to get the key SyKey(G)
            / Number of symmetric keys required is only one
            / Only the public-private key pair of the root cluster-head is needed
            / Encryption of SyKey(G) only happens once

Phase 2: Multicast session

if X is a group member and not the core
    it unicasts M encrypted with the key SyKey(G) to the core
    Core multicasts the encrypted message to all cluster members
    Each receiving cluster member decrypts with the key SyKey(G) to get M
else
    if X is a group member and is the core
        Core multicasts M encrypted with the key SyKey(G) to all cluster members
        Each receiving cluster member decrypts with the key SyKey(G) to get M
else
    if X is not a group member
        X unicasts to the core the message M encrypted with core's public key $PU_0$
                                        / X does not have the group key
        Core decrypts it with its private key $PR_0$ to get the message M
        Core multicasts M encrypted with the key SyKey(G) to all cluster members
        Each receiving cluster member decrypts with the key SyKey(G) to get M
```

Figure 7: Secured multicasting

The above multicast scheme uses only one group symmetric key (independent of the size of the group) and the public-private key pair of the root. This is an advantage, no doubt. However, problem like *man-in-the -middle-attack* may result if an intruder gets hold of the public key PU_0 of the root. In that case, it can easily decrypt E [PR_0, SyKey(G)] using PU_0 to get the group symmetric key. However, if the following group key distribution method as stated below is used in Phase 1, this problem due to intrusion can be avoided.

```
Phase 1: Distribution of group-symmetric key

1. Each cluster member $C_i^h$ (G) unicasts a join request Req to the core $C_0^h$
2. Core $C_0^h$ selects the cluster-symmetric key SyKey(G) and encrypts it with the public key $PU_i$ of $C_i^h$
            / each cluster-head knows the public keys of all other cluster-heads from TOI
3. Core $C_0^h$ unicasts the encrypted key SyKey(G) to each $C_i^h$ (G),
4. Each $C_i^h$ (G) decrypts with its private key $PR_i$ to get the key SyKey(G)
            / Number of symmetric keys required is only one
            / Number of public-private key pairs is the number of group members
            / Number of encryptions of SyKey(G) is the number of group members
```

In the above method of distribution, no intruder will have the knowledge of the private key of any joining node. So, it cannot get the group symmetric key anyway. However, this advantage comes at a cost; viz. now the required number of encryptions of SyKey(G) by the core is the number of group members whereas it is just one in the previous method, meaning thereby that multicasting based on this distribution will take more time to finish. In this context, it may be noted that multicasting inside a cluster [20] can be made secure in a similar way as above.

A question may arise like why not only public key-based approach is used for multicasting. It can be; however, it will be unnecessarily time consuming and it will not be truly multicasting. For example, consider a group of only five members, say m_1 to m_5. A source member say, m_1 wants to multicast a message M to other group members m_2 to m_5. So, first m_1 has to encrypt its message M separately with the respective public keys of the other group members and then it separately unicasts the four differently encrypted messages to the core to be sent to the four receivers m_2 to m_5; and then only the core can perform multiple-unicast to the receivers. First of all, there are too many encryptions and unicasts on behalf of the source causing it to be time consuming; secondly, it is not actually the idea of core-based multicasting. Whereas with symmetric key, source m_1 can just encrypt the message M once irrespective of the size of the group and unicasts it to the core. Only symmetric key distribution needs to be considered before multicast starts. Hence, the use of only a public key-based approach cannot be considered.

6 Conclusion

In this paper, we have considered the generalized form of a recently reported 2-layer non-DHT-based structured P2P network. In the generalized architecture efficiency of each data look-up, protocol remains the same as in the simpler initial version of the architecture. This has prompted us to consider the security aspect of the different communication protocols in the generalized architecture. To achieve it, public key-based approach has been used for the different look-up protocols (except for multicasting) because the required number of public-private key pairs is small. However, for multicasting among the cluster-heads, we have presented a hybrid approach using both symmetric key and public-private key ideas. It is shown that the number of symmetric keys required is just one independent of the size of the multicast group and only the public-private key pair of the root cluster-head is needed. To enhance security further while multicasting, we have shown that required number of public-private key pairs is just the number of cluster-heads.

As a continuation of our research, we are now working on secured communication in a federation of P2P architectures that consist of multiple Pyramid tree networks.

References

[1] L. Badis, M. Amad, D. Aîssani, K. Bedjguelal and A. Benkerrou, "ROUTIL: P2P Routing Protocol Based on Interest Links," 2016 International Conference on Advanced Aspects of Software Engineering (ICAASE), Constantine, pp. 1-5, 2016, doi: 10.1109/ ICAASE.2016.7843852.

[2] Tony A. Ballardie, "Core Based Tree Multicast Routing Architecture, Internet Engineering Task Force (IETF)," *RFC*, September 1997, 2201.

[3] Y. Chawathe, S. Ratnasamy, L. Breslau, N. Lanham, and S. Shenker, "Making Gnutella-Like P2P Systems Scalable," *Proc. ACM SIGCOMM*, Karlsruhe, Germany, pp. 407-418, August 25-29 2003.

[4] Shiping Chen, Baile Shi, Shigang Chen, and Ye Xia, "ACOM: Any-Source Capacity-Constrained Overlay Multicast in Non-DHT P2P Networks," *IEEE Tran. Parallel and Distributed Systems*, 18(9):1188-1201, Sept. 2007.

[5] Wen-Tsuen Chen, Chi-Hong Chao and Jeng-Long Chiang, "An Interested-Based Architecture for Peer-to-Peer Network Systems," 20th International Conference on Advanced Information Networking and Applications - (AINA'06), Vienna, 1:707-712, 2006, doi: 10.1109/AINA.2006.93.

[6] Jie Cheng and Ryder Donahue, "The Pirate Bay Torrent Analysis and Visualization," *IJCSET,* 3(2):38-42, Feb. 2013.

[7] Bidyut Gupta and Mohammad Mohsin, "Fault-Tolerance in Pyramid Tree Network Architecture," *J. Computer Systems Science and Engineering*, 10(3):164-172, July,1995.

[8] Bidyut Gupta, Nick Rahimi, Shahram Rahimi, and Ashraf Alyanbaawi, "Efficient Data Lookup in Non-DHT Based Low Diameter Structured P2P Network," *Proc. IEEE 15th Int. Conf. Industrial Informatics (IEEE INDIN)* , Emden, Germany, pp. 944-950, July 2017.

[9] B. Gupta, I. Roy, N. Rahimi, Z. Liu, and N. Debnath, "On Generalization of Non DHT-Based Pyramid Tree P2P Network Architecture," *IJCA,* 30(1):116-123, March 2023.

[10] M. Hai and Y. Tu, "A P2P E-Commerce Model Based on Interest Community," 2010 International Conference on Management of e-Commerce and e-Government, Chengdu, pp. 362-365, 2010, doi: 10.1109/ICMeCG. 2010.80.

[11] Mujtaba Khambatti, Kyung Ryu, and Partha Dasgupta, "Structuring Peer-to-Peer Networks Using Interest-Based Communities," Lecture Notes in Computer Science, 1st International Workshop, DBISP2P 2003, Berlin, pp. 48-63, September 2003.

[12] S. K. A. Khan and L. N. Tokarchuk, "Interest-Based Self Organization in Group-Structured P2P Networks," 2009 6th IEEE Consumer Communications and Networking Conference*,* Las Vegas, NV, pp. 1-5, 2009, doi: 10.1109/CCNC.2009.4784959.

[13] M. Kleis, E. K. Lua,, and X. Zhou, " Hierarchical Peer-to-Peer Networks using Lightweight SuperPeer Topologies," *Proc. IEEE Symp. Computers and Communications*, pp.143-148, 2005.

[14] D. Korzun and A. Gurtov, *Hierarchical Architectures in*

Structured Peer-to-Peer Overlay Networks, Peer-to-Peer Networking and Applications, Springer, pp. 1-37, March 2013

[15] Koushik Maddali, Indranil Roy, Swathi Kaluvakuri, Bidyut Gupta, Narayan Debnath, "Design of Broadcast Protocols for Non DHT-based Pyramid Tree P2P Architecture," *IJCA,* 28(4):193-203, December 2021.

[16] Anjila Neupane, Reshmi Mitra, Indranil Roy, Bidyut Gupta, Narayan Debnath, "Efficient and Secured Data Lookup Protocol using Public-Key and Digital Signature Authentication in RC-Based Hierarchical Structured P2P Network," *IJCA,* 30(1):140-150, June 2023.

[17] Z. Peng, Z. Duan, J.Jun Qi, Y. Cao, and E. Lv, "HP2P: A Hybrid Hierarchical P2P Network," *Proc. Intl. Conf. Digital Society*, pp. 18-28, 2007.

[18] N. Rahimi, K. Sinha, B. Gupta, and S. Rahimi, "LDEPTH: A Low Diameter Hierarchical P2P Network Architecture," *Proc. IEEE 14th Int. Conf. on Industrial Informatics (IEEE INDIN)*, Poitiers, France, pp. 832-837, July 2016.

[19] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker, "A Scalable Content-Addressable Network, CAN," *ACM*, 31(4):161-172, ACM, 2001.

[20] Arnold L. Rosenberg, "The Diogenes Approach to Testable Fault-Tolerant Arrays of Processors," *IEEE Tran. Computers*, c-32(10):902-910, Oct. 1983.

[21] A. Rowstron and P. Druschel, "Pastry: Scalable, Distributed Object Location and Routing for Large Scale Peer-to-Peer Systems." *Proc. FIP/ACM Intl. Conf. Distributed Systems Platforms (Middleware)*, pp. 329-350, 2001.

[22] Indranil Roy, Bidyut Gupta, Banafsheh Rekabdar, and Henry Hexmoor, "A Novel Approach Toward Designing A Non-DHT Based Structured P2P Network Architecture," EPiC Series in Computing, *Proceedings of 32nd Int. Conf. Computer Applications in Industry and Engineering*, 63:121-129, 2019.

[23] Indranil Roy, Swathi Kaluvakuri, Koushik Maddali, Abdullah Aydeger, Bidyut Gupta, and Narayan Debnath, "Capacity Constrained Broadcast and Multicast Protocols for Clusters in Pyramid Tree-based Structured P2P Network," *IJCA*, 28(3):12-18, Sept. 2021.

[24] Indranil Roy, Swathi Kaluvakuri, Koushik Maddali, Ziping Liu, and Bidyut Gupta, "Efficient Communication Protocols for Non DHT-Based Pyramid Tree P2P Architecture," (Invited paper), *WSEAS Transactions on Computers*, 20:108-125, July 2021.

[25] Indranil Roy, Koushik Maddali, Swathi Kaluvakuri, Banafsheh Rekabdar, Ziping Liu, Bidyut Gupta, Narayan Debnath, "Efficient Any Source Overlay Multicast in CRT-Based P2P Networks ─ A Capacity - Constrained Approach," *Proc. IEEE 17th Int. Conf. Industrial Informatics (IEEE INDIN)*, Helsinki, Finland, pp. 1351-1357, July 2019.

[26] Indranil Roy, Nick Rahimi, Ziping Liu, Bidyut Gupta, and Narayan Debnath*,* "On Generalization of Residue Class Based Pyramid Tree P2P Network Architecture," *IJCA,* 30(1):54-65, March 2023 Secured.

[27] H. Shen, G. Liu and L. Ward, "A Proximity-Aware Interest-Clustered P2P File Sharing System," *IEEE Transactions on Parallel and Distributed Systems*, 26(6):1509-1523, 1 June 2015, doi: 10.1109/TPDS.2014.2327033.

[28] K. Shuang, P Zhang, and S. Su, "Comb: A Resilient and Efficient Two-Hop Lookup Service for Distributed Communication System," *Security and Communication Networks*, 8(10):1890-1903, 2015.

[29] I. Stocia, R. Morris, D. Liben-Nowell, D. R. Karger, M. Kaashoek, F. Dabek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Protocol for Internet Applications," *IEEE/ACM Tran. Networking*, 11(1):17-32, Feb. 2003.

[30] Z. Tu, W. Jiang and J. Jia, "Hierarchical Hybrid DVE-P2P Networking Based on Interests Clustering," 2017 International Conference on Virtual Reality and Visualization (ICVRV), Zhengzhou, China, pp. 378-381, 2017, doi: 10.1109/ICVRV.2017.00087.

[31] M. Xu, S. Zhou, and J. Guan, "A New and Effective Hierarchical Overlay Structure for Peer-to-Peer Networks," *Computer Communications*, 34:862-874, 2011.

[32] M. Yang and Y. Yang, "An Efficient Hybrid Peer-to-Peer System for Distributed Data Sharing," *IEEE Trans. Computer*s, 59(9):1158-1171, Sep. 2010.

Nick Rahimi (photo not available) is the Director of Cyber Innovations Lab and an Assistant Professor at the School of Computing Sciences & Computer Engineering of the University of Southern Mississippi (USM). Dr. Rahimi obtained two Bachelor of Science degrees in Computer Software Engineering and Information Systems Technologies with a concentration in Cybersecurity and received his Master and Ph.D. degrees in Computer Science from Southern Illinois University (SIU). His research interests lie in the area of cybersecurity, blockchain, cryptography, internet anti-censorship, machine learning in cybersecurity, distributed systems, and decentralized networks. Prior to joining USM, Dr. Rahimi was a tenure track Assistant Professor at SIU for 2 years and Southeast Missouri State University (SEMO) for one year respectively. Moreover, he has over eight years of experience in industry as a software engineer and team leader.

Indranil Roy (photo not available) is an Assistant Professor in the Department of Computer Science at the Southeast Missouri State University. He received his MS and Ph.D. degrees in Computer Science from Southern Illinois University, Carbondale in 2018 and 2022, respectively. His current research interests include the design of architecture and communication protocols for structured peer-to-peer overlay networks, security in overlay networks, and Blockchain.

Ziping Liu (photo not available) received her PhD in Engineering Science from Southern Illinois University at Carbondale in 1999 and began her computing career at Motorola, where she developed software for mobile phones. Currently, she is a Professor of Computer Science at Southeast Missouri State University, where she has been teaching since 2001. Her research interests encompass a wide range of topics, including machine learning, cloud computing, secured software design, wireless ad hoc networks and sensor networks, distributed computing and game development.

Bidyut Gupta (photo not available) is currently a Professor of Computer Science at the School of Computing, Southern Illinois University at Carbondale. His research interests include fault tolerant distributed computing, design of P2P network architectures with low latency communication protocols, fog computing and its applications, and high latency networks. He is a Senior member of IEEE and ISCA.

Narayan C. Debnath (photo not available) is currently the Founding Dean of the School of Computing and Information Technology at Eastern International University, Vietnam. He is also serving as the Head of the Department of Software Engineering at Eastern International University, Vietnam. Formerly, Dr. Debnath served as a Full Professor of Computer Science at Winona State University, Minnesota, USA for 28 years, and the elected Chairperson of the Computer Science Department at Winona State University for 7 years. Dr. Debnath has been the Director of the International Society for Computers and their Applications (ISCA), USA since 2014. Professor Debnath made significant contributions in teaching, research, and services across the academic and professional communities. He has made original research contributions in software engineering, artificial intelligence and applications, and information science, technology, and engineering. He is an author or co-author of over 500 research paper publications in numerous refereed journals and conference proceedings in Computer Science, Information Science, Information Technology, System Sciences, Mathematics, and Electrical Engineering. He is also an author of over 15 books published by well-known international publishers including Elsevier, CRC, Wiley, Bentham Science, River Publishing, and Springer. Dr. Debnath has made numerous teaching, research and invited keynote presentations at various international conferences, industries, and teaching and research institutions in Africa, Asia, Australia, Europe, North America, and South America. He has been a visiting professor at universities in Argentina, China, India, Sudan, and Taiwan. He has been maintaining an active research and professional collaborations with many universities, faculty, scholars, professionals, and practitioners across the globe. Dr. Debnath is an active member of the IEEE, IEEE Computer Society, and a Senior Member of the International Society for Computers and their Applications (ISCA), USA.

An Evaluation Framework to Ensure the Quality of the Conceptual Models of Business Processes in a Biodiesel Plant

Narayan Debnath*
Eastern International University, VIENAM

Carlos Salgado†, Mario Peralta†, Daniel Riesco†, Lorena Baigorria†, and Germán Montejano†
Universidad Nacional de San Luis, ARGENTINA

Abstract

The objective of this work is to make a proposal to improve the quality of business processes in a biodiesel plant. As a first approximation, the analysis, and studies of the conceptual models of business processes that the company had were carried out, with the aim of being able to have a panoramic view of the current situation of the organization. A framework was applied to measure the quality of business process models, which provides a set of metrics and indicators to carry out said measurement. The objective of the frameworks is providing the organizations a means to help them to maintain objective and accurate information about the maintainability, understandability, comprehensibility, coupling and cohesion of the models, facilitating the evolution of the Business Processes of the companies involved in continuous improvement. It provides support to the management of BPs by facilitating early evaluation of certain quality properties of their models. The organizations benefit in two ways: (i) guaranteeing the understanding and dissemination of the BPs and their evolution without affecting their execution, (ii) reducing the effort necessary to change the models, this reduces the maintenance and improvement efforts. This framework is made up of two evaluation methods that face the same problem from two different approaches. One approach refers to the numerical and the other is closer to linguistic expressions similar to everyday language. Both methods provide important results to different areas of the business, giving the framework an added value when analyzing the BP conceptual models, since it allows organizations to choose the way to evaluate the models according to the characteristics that are desired to analyze the business models.

Key Words: Quality of conceptual models, business processes, workflow, evaluation method, and logics.

1 Introduction

The Business Process Modeling (BPM) is an essential step to achieve the objective of a Business Process (BP), as established by the BP definition, in order to obtain beneficial results for the stakeholders [14]. A BP model describes the activities involved in the business and how they are related to and how they interact with the necessary resources to achieve the objectives of the process [3]. From this point of view, the BPM is used to capture, document or redesign BPs [12].

The BP models present a global vision of the organization. This vision, allows a better understanding of the company´s dynamics and the relationships that occur within it and with its environment. This is the case both in the field that refers to customers and their suppliers and/or service providers. BPM is the technique par excellence to align the developments with the goals and objectives of organizations, since the models play a fundamental role in the specification of the BP. In the literature, you can find different conceptualizations of BP [7, 9, 15]. Therefore, the Workflow Management Coalition (WfMC) defines a BP as: *A set of two or more procedures or linked activities that collectively perform a business objective or a political goal, normally within the context of an organizational structure in which functional relationships and roles are defined* [17]. However, keeping in mind the various definitions of BPs, it can be said that, normally, a BP: (i) is associated with operational objectives and business relationships, (ii) may be contained entirely within an organizational unit or may encompass different organizations, (iii) has defined conditions that trigger its start, (iv) produce outputs defined at its completion, (v) may involve formal or relatively informal interactions among participants, and (vi) may consist of manual and/or automated activities. The conceptual model development represents a part of BP implantation. It is a key task of the first phase of BP life cycle. The users use the models as tools to easily understand the process that these models represent. Furthermore, they are the starting point when some changes or adaptations to new business needs are required for BPs. Therefore, the quality of these models is of vital importance to help improve the performance and evolution of the organization and do not become a risk factor. Under these

[*] School of Computing and Information Technology. Email: narayan.debnath@eiu.edu.vn.

[†] Departamento de Informática, Facultad de Ciencias Físico-Matemáticas y Naturales. Email: {csalgado, mperalta, driesco, flbaigor, gmonte}@unsl.edu.ar.

considerations, a framework is proposed to evaluate conceptual models of BP. The objective is to provide a means to organizations to help them study the quality of their BP models from the point of view of their **understandability** and **adaptability** to changes. Regarding these characteristics, the **understandability** allows the user to comprehend if the BP models are suitable and how to use them in particular tasks and conditions of use. It provides an indication of how easy it is to learn to read and interpret these models in order to understand the reality they are representing. The **adaptability** represents the capacity of the model to be modified effectively and efficiently due to evolutive, corrective or perfective needs. The modification of the model effectively and efficiently adapts without introducing defects or degrade their understandability.

Based on what was previously expressed, the application of a framework for the analysis and study of BP models from the perspective of the expected quality characteristics of a BP model of a medium company is presented. Said framework focuses on the use of continuous logic [4, 5] or fuzzy logic (FL) [18, 19], depending on the characteristics of the models and the processes that these models represent.

2 The Framework: F2BPM

The F2BPM framework is a means/tool that proposes the study and analysis of the BP models of an institution and/or organization. The main objective is to guide the development of such models by specifying requirements and evaluating quality characteristics. This framework is made up of three parts, each of which collaborates with the previous one. These parts are summarized as:

1. Apply the parser to the model, to determine its syntactic correctness.
2. Select and apply the evaluation method, according to the reality or the needs to be evaluated:
    a. Method based on continuous logic operators [1], or
    b. Method based on fuzzy logic [2].
3. Analyze results and generate reports and recommendations.

The Syntactic Analyzer checks the correct conformation of the BP model studied. At this stage, it is corroborated, for example, if the model meets or satisfies good practices or modeling guides for conceptual BP models [11]. These practices are summarized below:

G1: Minimize the amount of elements in a model, since its size has a negative impact on its understanding.

G2: Minimize the possible paths of each element, since the greater the number of inputs and outputs that an element has, the more difficult it is to understand.

G3: Indicate, as far as possible, a single starting element and a single final element in each process.

G4: Modeling in the most structured way possible by balancing the decision gates using the gates as parentheses: one to open in possible paths and another to close them to join them again.

G5: Avoid the use of OR gates, since the models containing only AND and XOR gates generally contain fewer errors.

G6: Use "verbal" type labels to define the actions of the tasks, for example "to analyze documentation" instead of "documentation analysis".

G7: Decompose the model if it has more than 50 elements, using, for example, sub-processes to make the general model more understandable.

Compliance with the modeling guidelines by the BP models means that the models are understandable and adaptable to the needs of the organizations, facilitating the task of the different actors that intervene in the process of BP conceptual modeling. Once the models are correct, we proceed to do the study and/or analysis through the use of one (or both) of the two alternatives provided by the F2BPM framework. Organizations can choose to work with: (i) operators of continuous logic, or (ii) to get closer to the natural and proper language of human beings through the use of *FL*, or some of its alternatives, or, if is not clear which of the alternatives is more appropriate for the particular situation. Both alternatives can be applied and then perform a comparative analysis of the results obtained.

The motivation of the method that works with operators of the continuous logic, [1], arises from the need of the organizations to have a means that allows them to represent their BP in an efficient way and that, in addition, allows them to communicate and interact with other processes. The objective of the method is to provide a means to help designers, analysts and developers involved in the definition and modeling of the BP of an organization to obtain models of quality processes. Throughout the phases of the method, the most relevant and frequent characteristics that BP conceptual models should satisfy are determined, grouped, and analyzed. These characteristics are reflected on a structure that will allow studying the degree to which the models satisfy them. In Figure 1 some desirable characteristics and sub-characteristics are detailed for all BP models.

1. Task/Activities
    1.1. Simples/Atomics
    1.2. Composed / Subprocesses
2. Synchronization points of the execution flow
    2.1. Decision Points
    2.2. Union Points
    2.3. Split Points in Parallel and / or Concurrent Execution
3. Events
    3.1. Start Events
    3.2. Intermediate Events
    3.3. Final Events
4. Participants / Actors
    4.1. Internals
        4.1.1. Number of Participants/Actors
        4.1.2. Communication between Participants / Actors

4.2. Externals
4.2.1. Number of Participants/Actors
5. Resources
5.1. Produced in the Process (Internals)
5.2. Externals

Figure 1: Requirement tree

In the next phase of method, elementary criteria are defined that will serve as measures of the degree to which the models evaluated satisfy the individual characteristics. To obtain the overall evaluation, these elementary criteria are combined until a single indicator of the overall satisfaction of the elementary characteristics is obtained, in order to finally carry out an analysis of the results obtained and outline the corresponding conclusions.

The second proposed method [2], is based on the fact that a FL system converts input variables (quantitative and qualitative) into linguistic variables through membership functions or fuzzy sets, which are evaluated by a set of fuzzy rules of the if-then type. Then, the outputs of the system become clear values (crisp) through a concretion process (*defuzzyfication*), which provides information for decision making. A FL system uses any type of information and processes it in a similar way as human thought. FL systems are adequate to treat qualitative, inaccurate, and uncertain information, which also allow dealing with complex processes, which makes it an interesting alternative for modeling decision-making problems. The diffuse control allows operating with vague or ambiguous concepts of the qualitative human reasoning, based on a mathematical support that allows extracting quantitative conclusions from a set of observations (premises) and qualitative rules (knowledge based).

When you have inaccurate and insufficient information, using statistical tools is not enough to obtain significant results. FL arises precisely to deal with this type of problem and to achieve an optimal solution. In this way, a combination between a FL system and the experience or knowledge that decision makers have is an excellent way to obtain good results. In the BP model development process the information about the business rules is usually imprecise or insufficient, which leads to the model being imprecise. Based on what was expressed, the use of FL in the evaluation of these models will allow, through the mechanisms provided by this logic, to evaluate imprecise and ambiguous situations produced in the construction of the models.

3 Evaluation of the PN Models of a Biodiesel Company

One of the fundamental aspects of any evaluation method is to show that it is of practical use. To achieve this goal, [13] presents a classification of three work proposals:

(i) Experimentation,
(ii) Cases of study,
(iii) Surveys.

For the practical validation of the method, and following the classification proposed in [13], a case study is presented in which the framework was applied to analyze the BP model of a local company. The decision to use case studies for the validation of the framework was due to the fact that, in general, there is no absolute control of the variables to be evaluated. This is because, in most cases, these variables depend on the particular reality that is being studied. For this reason, the application of case studies was considered more appropriate than the realization of experiments, in which it is necessary to have a greater control of the intervening variables, or the development of surveys, for which a certain history of application of the method and the opinion of those who used it should be available.

The main point was to improve the quality in a biodiesel plant in regards to the processes. But this brought with it an extra weight that had to be worked with a fuel that respects the environment, which does not contain sulfur, therefore it does not contribute to the greenhouse effect. It generates fewer emissions of polluting gases and substances that are harmful to health, such as carbon dioxide, carbon, soot, or benzene. These are some of the qualities of the fuel, but BP needed to be optimized in order to guarantee quality throughout the plant process. As a first task, a meeting was held with the directors of the company. There, they had access to the scant digitized documentation that the company had.

The general process is briefly described. There is an input of crude oil, to the degummed sub-process, which is rectified, removes the acidity, which eliminates fatty acids and phosphorus. To achieve this, they go through a trans-sterilization process to obtain biodiesel. In the middle of the process of obtaining biodiesel, it is possible to extract glycerin, that is also subjected to a process of separation of fatty acids that, to a lesser extent, have still remained from the trans-sterilization process. Glycerin is obtained in two varieties, a concentrated glycerin that is exported and another of lesser quality that is sold at a national level for the production of other products such as soaps among others.

Based on the above, the case study was the application of the framework for the evaluation of the BP model of the company, which aims to position itself successfully within the market. Thus, according to what the framework establishes, the BP model was analyzed to determine if it was syntactically correct through a parsing based on ALLOY [6, 8]. When checking this syntactic validity, it was necessary to prove if the model reliably represents the business logic. Part two of the framework was applied, in which the method to be used for the evaluation of the models is determined. In this stage, the method based on FL was chosen. The choice of this method was due to the use of FL [2], which allows a closer approach to the way of thinking of human beings; allows allowing the ambiguities that arise in terms of the interpretation of the different business rules that may arise. Below is a summary of the evaluation:

PHASE 1: Based on the BP model that the company owned (Figure 2) and using a set of defined metrics to measure the elements of a model, the data of the different components are collected. The amount of the different elements is counted as:

tasks, events, floodgates, among others. The information obtained is summarized in Table 1. It should be noted that, in this table, the metrics have been grouped according to the contribution they provide to the good practices associated with the BP model.

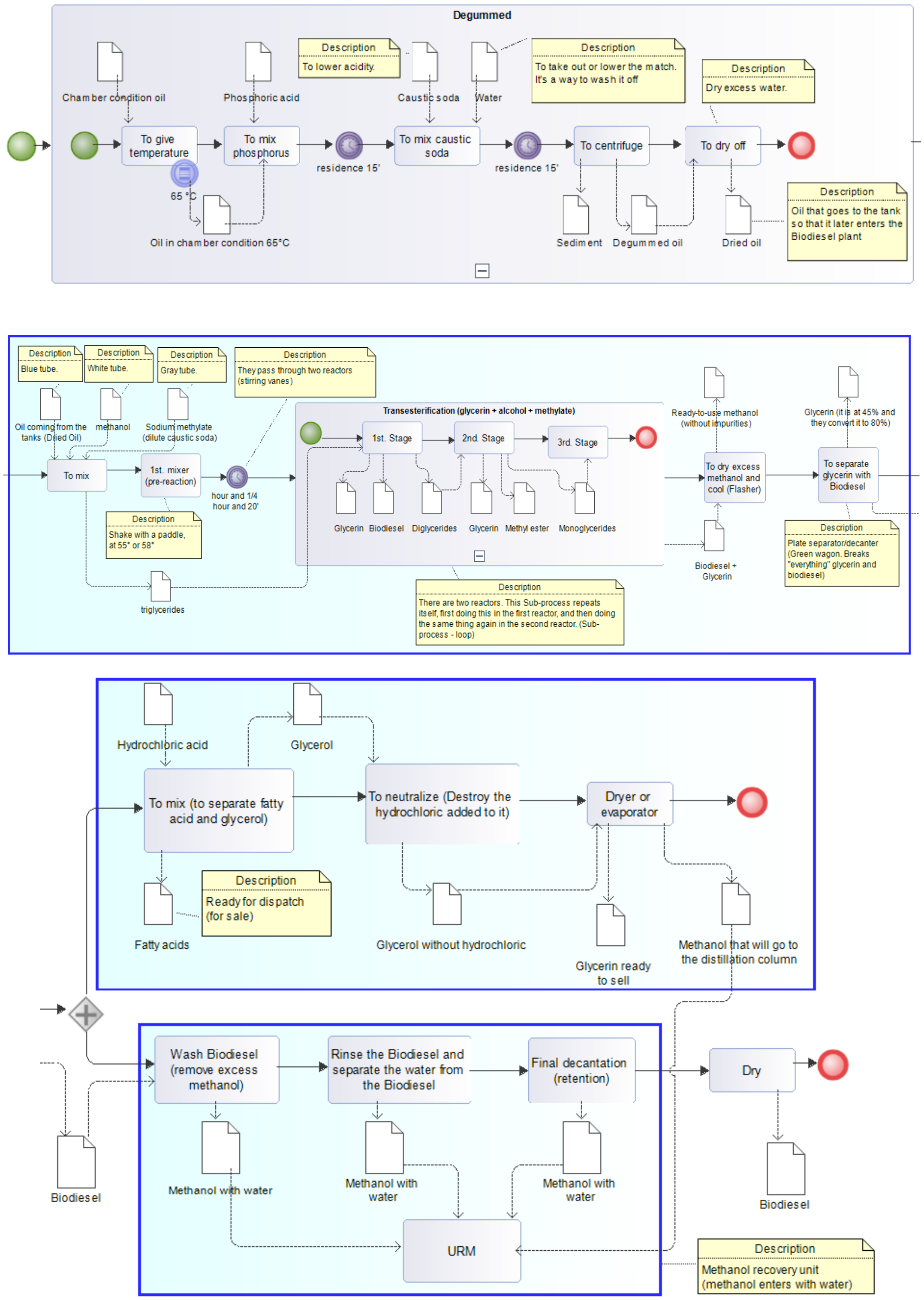

Degummed
Chamber condition oil
Phosphoric acid
Description
To lower acidity.
Caustic soda
Water
Description
To take out or lower the match. It's a way to wash it off
Description
Dry excess water.
To give temperature
To mix phosphorus
residence 15'
To mix caustic soda
residence 15'
To centrifuge
To dry off
65 °C
Oil in chamber condition 65°C
Sediment
Degummed oil
Dried oil
Description
Oil that goes to the tank so that it later enters the Biodiesel plant
Description
Blue tube.
Description
White tube.
Description
Gray tube.
Description
They pass through two reactors (stirring vanes)
Oil coming from the tanks (Dried Oil)
methanol
Sodium methylate (dilute caustic soda)
To mix
1st. mixer (pre-reaction)
hour and 1/4 hour and 20'
Description
Shake with a paddle, at 55° or 58°
triglycerides
Transesterification (glycerin + alcohol + methylate)
1st. Stage
2nd. Stage
3rd. Stage
Glycerin
Biodiesel
Diglycerides
Glycerin
Methyl ester
Monoglycerides
Description
There are two reactors. This Sub-process repeats itself, first doing this in the first reactor, and then doing the same thing again in the second reactor. (Sub-process - loop)
Ready-to-use methanol (without impurities)
To dry excess methanol and cool (Flasher)
Biodiesel + Glycerin
Glycerin (it is at 45% and they convert it to 80%)
To separate glycerin with Biodiesel
Description
Plate separator/decanter (Green wagon. Breaks "everything" glycerin and biodiesel)
Hydrochloric acid
Glycerol
To mix (to separate fatty acid and glycerol)
To neutralize (Destroy the hydrochloric added to it)
Dryer or evaporator
Fatty acids
Description
Ready for dispatch (for sale)
Glycerol without hydrochloric
Glycerin ready to sell
Methanol that will go to the distillation column
Wash Biodiesel (remove excess methanol)
Rinse the Biodiesel and separate the water from the Biodiesel
Final decantation (retention)
Dry
Biodiesel
Methanol with water
Methanol with water
Methanol with water
URM
Biodiesel
Description
Methanol recovery unit (methanol enters with water)

Figure 2: BP model

Table 1: Used metrics

TASKS/PROCESSES	GATES	EVENTS	PARTICIPANTS	RESOURSES
TNT = 18	TNGD = 0	TNSE = 2	NPI=12	NRI=8
TNCS = 0	TNGU = 0	TNIE = 4	NPE=2	NRE=23
	NPF = 1	TNEE = 4		

PHASE 2: The method proposes a set of variables of predefined inputs and outputs. This set can be extended, if necessary, by the members of the quality team involved in the process of quality evaluation of the models.

Of the suggested set of input/output variables, the following were taken for analysis. The choice of variables was limited to the metrics that allows the model to be addressed in terms of the components of most interest and that were the basis for the evaluation. In this sense, the following variables were taken:

INPUT: Number of Tasks/Activities of the Model, Number of Gates, Number of Processes, Match Start and Final Events, Number of Internal, Intermediate, and External Events; Numbers of both Internal and External Resources; Number of Internal and External Participants.

OUTPUT: Understanding of the Model, Maintainability of the Model, Coupling Level, Cohesion Level.

According to the phases of the method, once the variables have been defined, membership functions that indicate the degree of belonging of an element in a given universe must be established. The belongings functions that are used in this analysis are those proposed in the method presented in [2]. Figures 3 to 5 show these functions for the resources, events, and task/activities elements, respectively.

PHASE 3: At this stage, the decision to choose the knowledge base for the application of the method must be taken. The knowledge base contains the information associated with the domain of reality that is being studied.

Membership Functions	Function µ A(x)	Input Variables	Parameters

A1: Scarce	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x <= b
	0 if x > b
		Number of resources (internals and externals)	a = 5
			b = 19
			m = 12

A2: Several	0 if x < a
	(x-a)/(b-a) if a <= x <= b
	1 if b < x <= c
	(d-x)/(d-c) if c < x <= d
	0 if x > d
		Number of resources (internals and externals)	a = 14
			b = 20
			c = 27
			d = 33

A3: Many	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x <= b
	0 if x > b
		Number of resources (internals and externals)	a = 30
			b = 60
			m= 45

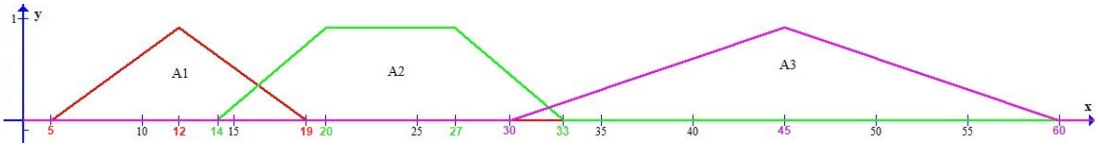


Figure 3: Membership functions for resources

Membership Functions	Function μ B(x)	Input Variables	Parameters

B1: Scarce	0 if x <= a
	$2*[(x-a)/(b-a)]^2$ if a < x <= m
	$1-2*[(x-b)/(b-a)]^2$ if m < x <= b
	1 if x > b
		Number of Events	a = 5
			b = 13
			m = 9

B2: Several	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x <= b
	0 if x > b
		Number of Events	a = 9
			b = 29
			m= 19

B3: Many	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x <= b
	0 if x > b
		Number of Events	a = 26
			b = 66
			m= 46

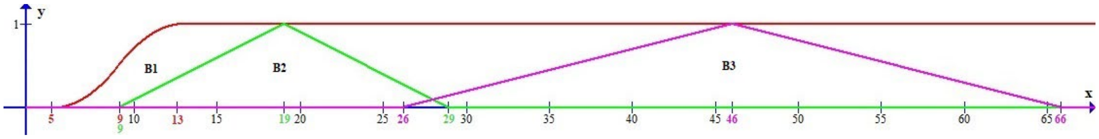


Figure 4: Membership functions for events

At this point, the linguistic rules must be defined that will serve to make decisions that, in turn will decide the way that the evaluator or who has to make the decisions must act. The rules follow the common sense of system behavior and are written in terms of the labels of membership functions. For the example that is studied there, a total of 30 rules are defined. Nevertheless, for the present example only the rules whose antecedents were calculated in the previous stage will be triggered. For example, for the variable resources: Scarce, Several, Many and for the variable Events: Scarce, Several, Many.

From the knowledge base, the subset of rules for the understandability feature is shown below.

Rule 1: IF (Several Resources) **and** (Several Events)
THEN Moderately Understandable
Rule 2: IF (Several Resources) **and** (Scarce Events)
THEN Mostly Understandable
Rule 3: IF (Many Resources) **and** (Several Events)
THEN Moderately Understandable
Rule 4: IF (Many Resources) **and** (Scarce Events)
THEN Mostly Understandable

PHASE 4: *Obtaining concrete values and system adjustments*: Until now, each of the four rules has been evaluated. The next step is to determine the fuzzy output by comparing the forces of all the rules that specify the same consequence, that is, the same output action.

At this stage, the ultimate goal is to find abrupt outputs. For this, each fuzzy output that was found in the previous stage of the evaluation rules, will modify their respective output membership function. The labels for these output functions refer to the understandability of the BP model, that is, they will be: Mostly Understandable, Moderately Understandable and Understandable. From the set of proposed output variables, this work is only concerned with model understandability (Figure 6) and comprehensibility (Figure 7).

A summary of the evaluation is shown below.

Input linguistic variables: Events, Resources.
Fuzzy Sets: Resources (x): Scarce, Several, Many (A1, A2, A3)
Events (y): Scarce, Several, Many (B1, B2, B3)
Output linguistic variables: Understandability of model
Fuzzy Sets: Understandability (w): Understandable, Moderately Understandable, Mostly Incomprehensible (D1, D2, D3).
Application of metrics to the model under study: **Events:** 10 and **Resources:** 31.

To Resources

$$\mu A1\ (x=31) = 0$$

$$\mu A2\ (x=31) = \frac{d-x}{d-c} = \frac{33-31}{33-27} = \frac{2}{6} = 0.33$$

Membership Functions	Function µ A(x)	Input Variables	Parameters

C1: Scarce	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x < b
	0 if x >= b
		Number of Task/Activity of Model
		Number of Processes, Number of Sub-processes
			a = 5
			b = 10
			m = 7

C2: Several	0 if x < a
	(x-a)/(b-a) if a < x <= b
	1 if b < x <= c
	(d-x)/(d-c) if c < x <= d
	0 if x > d
		Number of Task/Activity of Model
		Number of Processes, Number of Sub-processes
		Number Total of data Objects in the Model
			a = 7
			b = 20
			c = 45
			d = 75

C3: Many	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x < b
	0 if x >= b
		Number of Task/Activity of Model
		Number of Processes, Number of Sub-processes.
			a = 55
			b = 85
			m= 70

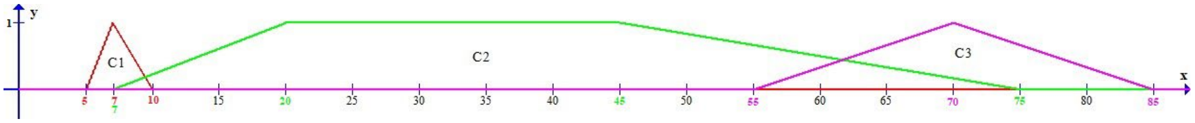


Figure 5: Membership functions for tasks/activities

Membership Functions	Function µ B(x)	Input Variables	Parameters

D1: Understandable	1 if b <= x <= c
	(d-x)/(d-c) if c < x <= d
	0 if x > d
		Number of Events	b = 0
			c = 20
			d = 30

D2: Moderately Understandable	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x <= b
	0 if x > b
		Number of Events	a = 25
			b = 55
			m= 40

D3: Mostly Understandable	0 if x <= a
	(x-a)/(m-a) if a < x <= m
	(b-x)/(b-m) if m < x <= b
	0 if x > b
		Number of Events	a = 50
			b = 85
			m= 70

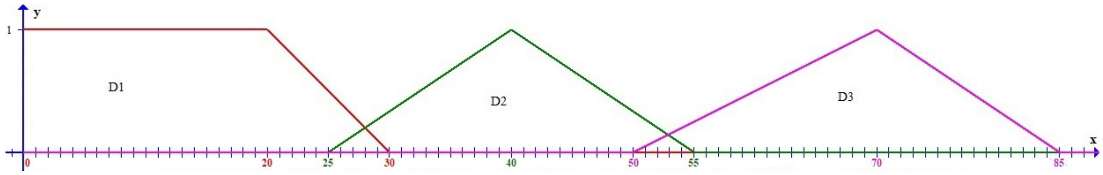


Figure 6: Membership functions for understandability

Membership Functions	Function µ B(x)		Input Variables	Parameters
D1: comprehensible	1	if b <= x <= c	Number of Events	b = 0
	(d-x)/(d-c)	if c < x <= d		c = 15
	0	if x > d		d = 30
D2: Moderately	0	if x <= a	Number of Events	a = 25
Comprehensible	(x-a)/(m-a)	if a < x <= m		b = 60
	(b-x)/(b-m)	if m < x <= b		m= 45
	0	if x > b		
D3: Mostly	0	if x <= a	Number of Events	a = 50
Comprehensible	(x-a)/(m-a)	if a < x <= m		b = 85
	(b-x)/(b-m)	if m < x <= b		m= 70
	0	if x > b		

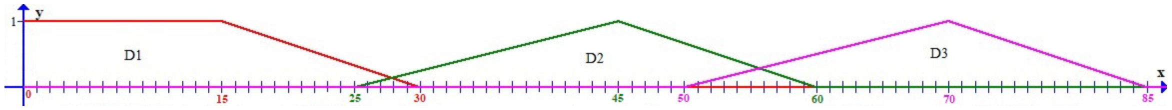


Figure 7: Membership functions for comprehensibility

$$\mu A3\ (x{=}31) = \frac{x-a}{m-a} = \frac{31-30}{45-30} = \frac{3}{7} = 0.06$$

To Events

$$\mu B1(y{=}10) = 1 - 2 * \left[\frac{y-b}{b-a}\right]^2 = 1 - 2 * \left[\frac{10-13}{13-5}\right]^2 = 0.72$$

$$\mu B2\ (y{=}10) = \frac{y-a}{m-a} = \frac{10-9}{19-10} = \frac{1}{9} = 0.11$$

$$\mu B3(y{=}10) = 0$$

Evaluation of Rules

Rule N° 1: $\mu A2(x)$ = 0.33; $\mu B2(y)$ = 0.11

Rule N° 2: $\mu A2(x)$ = 0.33; µB1(y) = 0.72

Rule N° 3: $\mu A3\ (x)$ = 0.06; $\mu B2(y)$ = 0.11

Rule N° 4: $\mu A3\ (x)$ = 0.06; $\mu B1(y)$ = 0.72

So far, each of the four rules has been evaluated. The next step is to determine the fuzzy output by comparing the forces of all the rules that specify the same consequence, that is, the same output action. In simple terms, if two or more rules try to affect the same output, the rule that is truer (of greater strength) will dominate. The method for the evaluation of rules used here is called MIN-MAX [10, 16], since it takes the minimum of the conditions to determine the strength of each rule and takes the stronger rule for each consequent, which determines the outputs.

The following values were obtained from the evaluation: $\mu B2(y)$=0,11 – $\mu A2(x)$=0,33 – $\mu A3(x)$=0,06 – $\mu A3(x)$=0,06.

Next, the Centroid of Gravity method is applied. Each output membership function is cut (lambda cut) at the level indicated by its respective output. The resulting cut membership functions are then combined to calculate their center of gravity:

$$\textbf{Output} = \frac{\sum_{i=1}^{n} x_i \mu_c(x_i)}{\sum_{i=1}^{n} \mu(x_i)}$$

Output

$$= \frac{75*0{,}11+255*0{,}33+525*0{,}06}{1{,}2+1{,}98+0{,}06} = \frac{8{,}25+84{,}15+31{,}5}{3{,}24} = \frac{123{,}9}{3{,}24} = 38{,}24$$

Similarly, the evaluation is carried out for the comprehensibility characteristic of the model.

A summary of the evaluation is shown below.

Input linguistic variables: Events, Resources.
Fuzzy Sets: Resources (x): Scarce, Several, Many (A1, A2, A3)
Events (y): Scarce, Several, Many (B1, B2, B3)
Output linguistic variables: Comprehensible of model
Fuzzy Sets: Comprehensible (w): Comprehensible,
Moderately comprehensible, Mostly comprehensible
(D1, D2, D3).

Application of metrics to the model under study: **Events:** 10 and **Resources:** 31.

To Resources

$$\mu A1\ (x=31) = 0$$

$$\mu A2\ (x=31) = \frac{d-x}{d-c} = \frac{33-31}{33-27} = \frac{2}{6} = 0.33$$

$$\mu A3\ (x=31) = \frac{x-a}{m-a} = \frac{31-30}{45-30} = \frac{3}{7} = 0.06$$

To Events

$$\mu B1(y=10) = 1 - 2 * \left[\frac{y-b}{b-a}\right]^2 = 1 - 2 * \left[\frac{10-13}{13-5}\right]^2 = 0.72$$

$$\mu B2\ (y=10) = \frac{y-a}{m-a} = \frac{10-9}{19-10} = \frac{1}{9} = 0.11$$

$$\mu B3(y=10) = 0$$

Evaluation of Rules

***Rule N° 1**:* $\mu A2(x) = 0.33$; $\mu B2(y) = 0.11$

***Rule N° 2**:* $\mu A2(x) = 0.33$; $\mu B1(y) = 0.72$

***Rule N° 3**:* $\mu A3\ (x) = 0{,}06$; $\mu B2(y) = 0.11$

***Rule N° 4**:* $\mu A3\ (x) = 0{,}06$; $\mu B1(y) = 0.72$

For the components, processes, threads, and tasks that make up the analyzed model, the output is moderately understandable and comprehensible.

PHASE 5: *Analysis and Documentation of the Obtained Results*: This stage corresponds to the final stage of the method. In it, an analysis and comparison of the results obtained in the evaluation of the models with respect to the preferences of the users, obtained in the application of the method, must be carried out. In addition, the evaluation process and the obtained results must be documented, so that the documentation serves as a reference and history of the evolution of the studied BP model in future evaluations of models. This documentation can serve as a point of reference and comparison at the evaluation of new models and BPs. This phase deals with activities of analysis and comparison of the preferences of quality and the obtained results. Based on the established goals, and the point of view of those interested in the models and BPs to be evaluated, this stage culminates with the conclusions and recommendations of the case.

Both framework methods define a phase of analysis of results. This stage is one of the most relevant activities of the framework. Therefore, it is extremely useful to have the information gathered during the application of the method selected for the evaluation collected on structures and representations that are clear to read and interpret. The method proposes a standard form that should be completed once the evaluation of the models has been carried out. The form allows, among other things, what membership functions were used; if they were defined by the evaluating group or if others previously defined and stored in a repository were used. There is also information about the models evaluated, and the analysis of the results obtained. In addition, data of the models, the evaluators are recorded and, if there are previous evaluations, a reference to them is included. These references serve as a point of contrast to analyze and evaluate the evolution of the models.

Finally, a field is included where a report of the analysis of the results can be presented. For reasons of space, the structure of the form is not shown in this paper. At this point, when analyzing the results of the framework application, emphasis was placed on the understandability of the results delivered and presented in the forms, and the perception of the different actors involved in the modeling process. For the case study, we worked with a group of 20 people, including administrative staff, technicians, quality staff, analysts, and designers. From the calculation of the output made in the previous point it can be seen that the model studied is moderately understabdable and comprehensible.

4 Conclusions

Continuous improvement is a fundamental tool for all companies because it allows renewing or improving their BP. This implies a constant updating that makes the organizations more efficient and competitive. The BP model is the basis for better understanding the operation of an organization, documenting, and publishing the processes seeking standardization in the organization, achieving greater efficiency in the operation, and integrating solutions in service-oriented architectures. These characteristics give the organization a valuable tool to stay at a competitive level. Thus, the BP models are fundamental when analyzing the correctness and quality of the processes that they model.

From this point of view, the use of the framework is proposed with the possibility of applying any of the two methods that compose it. The first of the methods pursues the objective of providing organizations with a means to help them maintain objective information about the maintainability of the models. This facilitates the evolution of the BPs of the companies that constantly evaluate their processes to be involved in continuous improvement. In addition, it provides support to the management of BPs by facilitating the early evaluation of certain quality properties of their models. With this, the organizations benefit in two ways: (i) guaranteeing the understanding and dissemination of the BPs and their evolution without affecting their execution; (ii) and reducing the effort necessary to change the models with the consequent reduction of maintenance and improvement efforts.

However, when developing BPMs, information about the business rules that must be represented and modeled is often imprecise or insufficient, leading to inaccurate models. From this perspective, FL provides mechanisms to analyze and simulate human reasoning. Therefore, the use of fuzzy logic in the evaluation of BPM allowed, through the mechanisms provided by this logic, to evaluate those imprecise and ambiguous situations produced in the construction of this model.

Under these considerations, the BPs of a biodiesel producing company were analyzed, looking for decision-making points that would limit losses, waste, and plant shutdowns for different reasons. This paper presents one of the tasks that were carried out in order to understand and control the biodiesel production process. Among the points to analyze that were detected at the beginning were: 1. Waiting in production for complete deposits, 2. Idle times of the employees, 3. Shutdown of the plant due to waiting time for: a) Raw materials; b) Withdrawal of Fuel, 4. Management estimates.

In the BP models studied, they reflected that plant stoppages occurred repeatedly, because the tanks were full and the biodiesel was not withdrawn by the customers. This caused lost time for employees with the expenses that were incurred and final products in warehouses without the possibility of having movements. Initially, it led to expanding the capacity of the storage tanks to 1,000,000 liters, what relaxed the stoppage of the plant from the perspective of stock management. But there was still a problem with the logistics of the clients in the withdrawal of the merchandise. Regarding the purchase of the raw material, it was possible to see that there was a gap between the communication of the internal actors of the biodiesel plant. An excess of internal messages with unnecessary waiting time for administrative order, that they produced a plant stoppage every 45 days, due to a problem of an internal logistical/administrative nature. This ends up being solved by setting a minimum stock alert that anticipated purchases.

References

[1] N. Debnath, C. Salgado, M. Peralta, M. Berón, D. Riesco, and G. Montejano, "MEBPCM: A Method for Evaluating Business Process Conceptual Models, A Study Case," Presented at the 9th ITNG, Las Vegas, Nevada, USA, 2012.

[2] N. Debnath, C. Salgado, M. Peralta, D. Riesco, L. Baigorria, and G. Montejano, "A Fuzzy Logic-Based Method to Evaluate the Quality of Business Process Models," CATA 2017, 2017.

[3] T. Dufresne and J. Martin, "Process Modeling for e-Business," Spring 2003, INFS 770 - Methods for Informations Systems Engineering: Knowledge Management and E-Business, George Mason University, 2003.

[4] J. J. Dujmovic, "Quantitative Methods for Software Evaluation," Lecture Notes, Graduate Software Engineering Program. Universidad Nacional de San Luis, San Luis, Argentina., 1998.

[5] J. J. Dujmovic, "A Method for Evaluation and Selection of Complex Hardware and Software Systems," *The 22nd International Conference for the Resource Management and Performance Evaluation of Enterprise Computing Systems. CMG96 Proceedings,* 1:368-378, 1996.

[6] R. Gheyi, T. Massoni, and P. Borba, "Formally Introducing Alloy Idioms," Brazilian Symposium on Formal Methods., pp. 22-37, 2007.

[7] Z. Irani, V. Hlupic, and G. M. Giaglis, "Business Process Re-Engineering: a Modeling Perspective," *International Journal of Flexible Manufacturing Systems,* Need volume number, 13:99-104., 2001, https://doi.org/10.1023/A:1011122916139.

[8] D. Jackson, *Software Abstractions: Logic, Language, and Analysis*, MIT Press, Cambridge, 2006.

[9] C. Jiménez, L. Farías, and F. Pinto, "Análisis de Modelos de Procesos de Negocios en Relación a la Dimensión Informática," *Revista Electrónica del DIICC.* http://www.inf.udec.cl/revista/ediciones/edicion9/cjimenez.pdf, 2004.

[10] D. P. Madau and L. A. Feldkamp, "Influence Value Defuzzification Method," *Proceedings of IEEE 5th International Fuzzy Systems,* New Orleans, LA, USA, 3:1819-1824, 1996, doi: 10.1109/FUZZY.1996.552647.

[11] J. Mendling, H. Reijers, and W. van der Aalst, "Seven Process Modeling Guidelines," *Information & Software Technology,* 52(2):127-136, 2010.

[12] B. Mora, F. Ruiz, F. García, and M. Piattini, "Experiencia en Transformación de Modelos de Procesos de Negocios desde BPMN a XPDL.," IDEAS, 2007.

[13] C. Robson, *Real World Research: A Resource for Social Scientists and Practitioners-Researchers*: Blacwell, 1993.

[14] A. Sharp and P. McDermott, "Workflow Modeling: Tools for Process Improvement and Application Development," London: Artech House, 2001.

[15] G. Sparks, "An Introduction to UML," The Business Process Model. Sparx Systems, www.sparxsystems.com.au, 2000.

[16] W. van Leekwijck and E. E. Kerre, "Defuzzification: Criteria and Classification, Fuzzy Sets and Systems," pp. 159-178, 1999.

[17] WfMC, "Terminology & Glosary," Workflow Management Coalition, Document Number: WFMC-TC-1011. Document Status: Issue 3.0, 1999.

[18] L. A. Zadeh, "Fuzzy Sets," *Information and Control,* 8:338-353, 1965.

[19] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex System," *IEEE Transaction on System Man and Cybernetics,* 1:28-44, 1973

.

Narayan Debnath (photo not available) earned a Doctor of Science (D.Sc.) degree in Physics. Narayan C. Debnath is currently the Founding Dean of the School of Computing and Information Technology at Eastern International University, Vietnam. Dr. Debnath has been the Director of the International Society for Computers and their Applications (ISCA) since 2014. Formerly, Dr. Debnath served as a Full Professor of Computer Science at Winona State University, Minnesota, USA for 28 years (1989-2017). Dr. Debnath has been an active member of the ACM, IEEE Computer Society, Arab Computer Society and a senior member of the ISCA.

Carlos Celular (photo not available) earned a Master's degree in Software Engineering. He currently works as Adjunct Professor of the Department of Informatics, of the Faculty of Physical-Mathematical and Natural Sciences, of the National University of San Luis, Argentina. He is a Professor in undergraduate, graduate, and postgraduate career. He is the director of the undergraduate degree in Web Technology and co-director of the Master's degree in Software Quality careers. He is a member of different career committees of the Computer Science Department and postgraduate professor in the Master's Degree in Software Quality, Master's Degree in Software Engineering, Specialization in Software Engineering. He is a specialist in quality models and evaluation methods in software engineering processes and comprehensive professional in the area of analysis, design, development, and implementation of systems applied to business processes. He has extensive knowledge and expertise related to quality, both at the level of conceptual models of business processes, process workflows, as well as in terms of product quality. He is experienced in evaluating and comparing desirable features in products such as web and mobile applications. His skills and abilities are related to the use of metrics and quality indicators, optimization of business processes, with extensive expertise in business benchmarking, workflow models, planning and services including consulting in engineering and reengineering of business processes and development of business logistics. He is a member of the Software Engineering Quality Laboratory, and a member of various research projects in Software Engineering from 2002 to the present.

Mario Peralta (photo not available) Obtuvo un título de Especialista en Ingeniería de Software. Actualmente se desempeña como Profesor Adjunto del Departamento de Informática, de la Facultad de Ciencias Físico-Matemáticas y Naturales, de la Universidad Nacional de San Luis, Argentina. Profesor en carreras de pregrado, grado y posgrado. Codirector y profesor de la carrera de posgrado Especialización en Ingeniería de Software. Su área de experticia se centra en la Gestión de Procesos de Negocio y Procesos Worflow, en el análisis, diseño, desarrollo e implantación de sistemas aplicados a procesos de negocio. Conocimientos y experiencia en análisis de calidad de modelos de procesos de negocio, flujos de trabajo de procesos y de productos software. Habilidades y capacidades en la utilización de métricas e indicadores de calidad, optimización de procesos de negocio, como así también en la evaluación comparativa de negocios, modelos workflow, planificaciones y servicios. Experiencia en la administración de herramientas de gestión de procesos workflow corporativos. Es integrante de diversos proyectos de investigación en Ingeniería de Software desde 2002 y hasta la actualidad. Actualmente se desempeña como Director del Departamento de Informática, de la Facultad de Ciencias Físico-Matemáticas y Naturales, de la Universidad Nacional de San Luis.

He earned a Software Engineering Specialist degree. He currently works as Adjunct Professor of the Department of Informatics, of the Faculty of Physical-Mathematical and Natural Sciences, of the National University of San Luis, Argentina. He teaches undergraduate, graduate, and postgraduate careers. He is co-director and professor of the postgraduate course Specialization in Software Engineering. His area of expertise focuses on Business Process Management and Workflow Processes, on the analysis, design, development, and implementation of systems applied to business processes. He has knowledge and experience in quality analysis of business process models, process workflows and software products. His skills and abilities include the use of metrics and quality indicators, optimization of business processes, as well as in the comparative evaluation of businesses, workflow models, planning and services. He is experienced in the administration of corporate workflow process management tools. He is a member of various research projects in Software Engineering from 2002 to the present. He currently works as Director of the Department of Informatics, of the Faculty of Physical-Mathematical and Natural Sciences, of the National University of San Luis.

Daniel Riesco (photo not available) Obtained a PhD in Software Engineering from the University of Vigo, Spain. He currently works as an Associate Professor of the Department of Informatics, of the Faculty of Physical-Mathematical and Natural Sciences, of the National University of San Luis, Argentina. He is a professor in undergraduate and postgraduate courses. He is Director of the graduate programs in Master's in Software Engineering and PhD in Computer Engineering. He has proven experience in Software Engineering and Information Systems, in particular areas of Management and Modeling of Business Processes, Software Quality, Software Development Processes, Model Driven Engineering, Software Architecture, Cloud Computing, Web Services, and other web technologies. He is a member and director of various research projects in Software Engineering.

Lorena Baigorria (photo not available) earned a Master's degree in Software Engineering. She currently works as Adjunct Professor of the Department of Informatics, of the Faculty of Physical-Mathematical and Natural Sciences, of the National University of San Luis, Argentina. She teaches undergraduate, graduate, and postgraduate courses. She is director of the Computer Science Engineering degree career and co-director of the Master's in Software Engineering and a member of different career committees of the Computer Science Department. She is a postgraduate professor in the Master's Degree in Software Quality, Master's Degree in Software Engineering, Specialization in Software Engineering. Her research areas focus on high quality software development management including analyzing usability improvements and software quality applying new technologies as orientation to aspects, mobile developments, cloud computing, web services. She is a member of various research projects in Software Engineering from 2002 to the present.

Germán Montejano (photo not available) obtained a PhD in Software Engineering from the University of Vigo, Spain. He currently works as an Associate Professor of the Department of Informatics, of the Faculty of Physical-Mathematical and Natural Sciences, of the National University of San Luis, Argentina. He is a professor in undergraduate and postgraduate courses. He is the director of the postgraduate courses Master's in Software Quality and Specialization in Software Engineering and co-director of the PhD in Computer Engineering. His area of expertise is focused on Software Engineering, Software Project Management, Financial Evaluation of Software Projects, Information Security, Cyber Security, Cyber Defense, Cloud Computing, Internet of Things. He is an experienced Professor with a demonstrated history of working in the higher education industry. He has a strong education professional skilled in Computer Science, Java, HTML, Linux, and Algorithms. He is a member and line director of various research projects in Software Engineering.

Hybrid Remote Work Models in Project-Organized Small and Medium-Sized IT Companies

Tomaž Kokot[*]
Alma Mater Europaea - ECM, SLOVENIA

Abstract

Remote work used to be more of an exception until today, mainly used around the world, especially in the field of ICT. In Slovenia there was not so much practice and discussion about this way of working until recently. Unintentionally, this way of working has spread due to the COVID-19 pandemic. It has not only changed the characteristics of workplaces, but also the attitude of employees towards this way of working. The employees tried their hand at working remotely, and many of them perceived the potential and the desire to continue doing so. Namely, many advantages were shown, among which the flexibility of time and, of course, the long-desired balance between professional and private life stand out. Businesses have also seen many benefits, such as a reduction in overall operating costs, a way to reduce turnover and a way to increase productivity. Among other things, project management is also reflected through methodological diversity, which in our study is reflected through an empirical perspective, since in the research we used a triangulation approach between qualitative and quantitative data. A survey was conducted among 100 respondents with the help of a questionnaire, with which we gained insight into the positive and negative experiences of remote work in small and medium-oriented (SMITCs) IT companies. The obtained results were subsequently analyzed with the help of decision trees, and based on these results these questions were pre-debated by the focus group. The research showed that employees have mostly positive experiences when working remotely, which affects their motivation, satisfaction and, of course, productivity. Considering that in this case it was remote work that was not originally planned, the company will definitely need some improvements for this type of work. The question however remains if the ICT sector will further support remote work gained in recent period or will use a hybrid model. Currently, the hybrid model provides the opportunity to choose the best advantages of remote work and the classic way of working and achieve a rhythm that will satisfy both the management and the employees. What the future of work in the field of ICT will be, however, is still impossible to accurately predict at the moment, despite the prospects.

Key Words: Hybrid model, telecommuting, productivity, IT companies.

1 Introduction

Working remotely is not a new way of working. The beginnings of trying to do so began at the early part of the 21st century. At that time, remote work represented an alternative form of work, which only here and there supplemented work in the classic way due to additional needs. Later telecommuting took place only sporadically, mainly to support work-life balance and thus support flexibility for employees to balance their role at work and outside of it (Afrianty, Artatanaya, and Burgess) [1]. However, this changed when communication technology (hereafter ICT) began to rapidly develop. The greater the progress of ICT, the more telecommuting was available, or it became more acceptable in the organizational context and employers chose this way of working more often. In those days, it was rarely intended for an employee to telecommute full-time (Peek) [7].

Remote work has started to become more common in certain fields world-wide. It was also becoming more and more acceptable, but still, many did not think of trying this way of working. Technology certainly made it possible to do some work both from the office where the company is located, as well as from anywhere in the world, especially in the field of work where a service can be performed or provided online (Arruda) [2]. When the COVID-19 pandemic began to spread around the world and the measures became more and more strict, employees were actually forced to work from home. Employers began to seriously think about eliminating physical workplaces. Moreover, as Dayaram and Burgess [5] explain, there has been another noticeable change, namely the introduction of remote work among professions, where this was not possible previously or they did not even think about it. While reporting on how employers and employees find themselves during remote work, we were able to get some insights about how some employers are thinking about the complete digitization of work. Employees noticed benefits with work that the employees would perform solely and only as remote work, while the company would not have any physical premises at all (Shaner) [8].

Barrero, Bloom and Davis [3] state that COVID-19 has launched a massive social experiment. The data is an excellent indicator for such a claim. In 2020, from April to December, Americans worked about half of their paid work hours from home. Before the pandemic, these hours were by an average of five percent. Despite the fact that some are completely against remote work because they are worried about the lack of personal relationships, others are convinced that the adjustment has been incredible and that more work will be done from home in the future, a kind of intermediate

[*] PhD student. Email: Tomaz.Kokot@AlmaMater.si.

view is also expressed. It is recognized that some things actually work very well if they are done virtually only. In their contribution, the mentioned authors explained that they are developing systematic evidence on whether remote work will survive and why. Their research was based on data on the frequency of telecommuting, the possibility of switching to such a way of working and data on the well-being of employees working remotely and the impact of telecommuting on productivity, costs and other related factors. Their conclusions, which can be an introduction to our research, showed that remote work will remain, long after the end of the pandemic. Employees in America are expected to do up to 20 percent of their work from home, which is four times more than before. They also found that the willingness of some employees to telecommute has increased to the point that they are willing to work for much less pay in exchange for the opportunity to telecommute at least two or three days a week. However, another side of the work appears, namely the impact on the economy in terms of challenges for urban areas. By reducing transportation to work, employees will also reduce some other needs, such as less shopping, less personal services, less entertainment, which means a reduction in total consumer spending (Barrero, Bloom, and Davis) [3]. The question that arises here is whether the increase in productivity will be such that other, consequential effects on the economy can be neglected. In fact, the increase in productivity is often possible precisely because of time savings and cost reduction, which may not mean anything for ordinary economic statistics.

The restrictions caused by COVID-19 have contributed to the fact that there is an increasing need for alternative work methods, which include remote work and work defined as a hybrid model. The hybrid work model is not a newly discovered work method, but due to recent events, the need for such a model is increasing. It is therefore a model also known as a mixed system, which usually appears when it is necessary to balance two types of requirements. In the case of this model, we are talking about combining physical work arrangements (work performed at the employer's location, i.e., in the office) and remote work (Cook, Mor, and Santos) [4]. The hybrid work model thus combines remote work, where the essential advantages are flexibility at work, lower labor costs and greater employee satisfaction, and the advantages of the traditional work system, such as personal cooperation and better culture in the work organization. The need for remote work is actually growing faster and faster, which offers researchers the opportunity to more thoroughly investigate one and the other work model (Iqbal, Evgenevich Barykin, and Khalid) [6].

In our study, we discuss the possibilities of both systems, namely, we studied the advantages of remote work for employees in the IT SMITCs companies, and at the same time we checked whether the use of a hybrid model would be more appropriate in this area. Due to the COVID-19 pandemic, the employees experienced remote work, this was actually an accelerating factor that now allows us to check what the response was. Changes are needed for the future, based on digitization, which will have to be restructured so that the company does not become obsolete with its way of doing business. As a result, there are more and more opportunities for hybrid work arrangements in many areas. In the first part, we present the evaluation of the research problem by means of connecting the key theoretical findings of the studied field.

Project management is a complex process that involves many important competencies. Its role in the case of research, however, has changed significantly recently. Sometimes research was based on the assessment of factors and appropriateness of use and contribution to knowledge and expertise. Today, the use of project management in research is based on the integration of qualitative and quantitative studies and thus a comprehensive review of the practical results obtained. The empirical part first presents the analysis of the survey questionnaire in which we obtained the preliminary results for further investigation using the decision trees. This was followed by a presentation of the questions and outputs to the focus group to debate the opinions and positions on the mentioned questions.

1.1 Purpose and Goals

The purpose of the study is to expand the insight into remote work and the possibilities of using the hybrid model in the case of IT SMITCs companies. Based on the findings, we evaluated the problem presented and the goals of the research were aimed at:

- to study methods based on project management methodology;
- presentation of the results obtained with the help of a survey questionnaire, where employees in the IT field evaluated the positive and negative aspects of telecommuting factors;
- presentation of employee productivity modeling using decision trees and analysis of obtained results and presentation of priority factors related to employee productivity;
- presentation of the questions asked obtained on the basis of the results of decision trees and the use and analysis of the selected measuring instrument to gather opinions and positions with the help of discussion;
- analyzing the obtained results and their interpretation and
- presentation of an understanding of the advantages of using a standard work system, remote work and hybrid work in the case of IT companies.

With the aforementioned research we determined what kind of work is most suitable for employees in IT SMITCs companies. It has the most advantages, standard work in the office, remote work, based on research in the past almost two years, or for companies in this field in the future, to consider a hybrid model. In the study, we also sought answers to the following hypotheses:

H1: The perception of remote work depends on the field of work.

H2: The positive effects of telecommuting in the case of the IT sector can affect productivity, and in the case of the hybrid model, these could increase even more.

H3 Employees see many advantages in remote work, but the standard form of work is also close to them, which means that the hybrid model could be the most effective form of work in the IT sector.

1.2 Methods

We conducted the survey on the experience of working remotely when the situation during the pandemic had calmed down to such an extent that employees returned to their jobs in the office. The questionnaire was designed with the help of theoretical knowledge and past research on telework, based on factors related to the impact of telework on productivity. One hundred IT respondents evaluated general claims about telecommuting, positive and negative factors in telecommuting, and factors directly related to productivity in telecommuting. The respondents gave their evaluations using a measurement scale, with grades from 1 to 5 (1-do not agree at all, 5-completely agree), so that they evaluated the individual statements in each group. We also checked demographic data such as: gender, age, completed education, employment status, number of years with the current employer and personality type. We analyzed the obtained data of the survey questionnaire according to the evaluation of the respondents. We then evaluate results with the decision trees method systematically based on all possible outputs. The decision tree method was chosen as a method that allows a symbolic representation of patterns, which we can use and reinterpret. The dependent variable was shifted, for example when stated to telecommuting productivity which was derived from a set of statements in the survey questionnaire, we used other variables such as demographic data etc. as independent variables.

We analyzed the data sets obtained with the help of decision trees and selected the 10 most important findings, which highlight the essential findings of the research and their connection with the positive and negative aspects of remote work. We analyzed, presented and interpreted the collected data accordingly. On the basis of these results, we have posed a meaningful question in light of the findings, which would help us to explain the obtained data more deeply. We continued with a qualitative research in which we used the focus group method, where 7 participants were informally interviewed in a group. We chose this method in order to obtain general information about the background of the researched topic. We continued with the presentation of the collected results of the focus groups, and for an easier and more in-depth understanding of the results. We also used some quotes from the focus group participants. Our workflow progress is illustrated by Figure 1.

2 Research

One hundred respondents rated the statements in the questionnaire as follows:

- The majority of respondents completely agreed (rating 5) with the fact that working remotely saves time (e.g., no driving to work) and consequently also costs, and that such a method of work requires a high level of digital competence or literacy.
- The majority of respondents agreed (rating 4) that: when working remotely, free choice of time or flexibility of working hours is important (to do the work when they think they will be most productive); that they have a better balance between their professional and private lives when working remotely; that telecommuting is a challenge; that they are more creative when working remotely; that when working
- remotely there are no usual distractions that are present when working at a traditional workplace (e.g. distractions from other employees); that telecommuting requires a certain level of responsibility towards work; that their productivity is higher when they work from home; that remote work has more advantages than traditional work in an office; that when working remotely, greater productivity depends on the level of
- motivation and their satisfaction (the higher both factors are, the higher the productivity) and that an additional

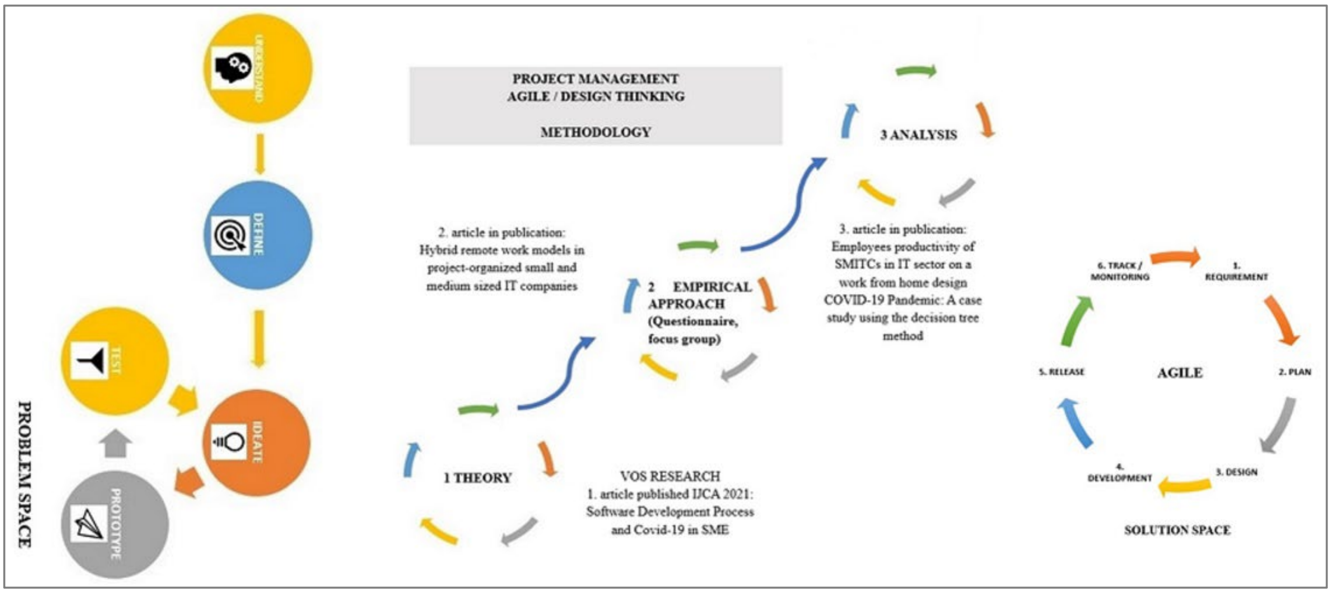


Figure 1: The workflow in our research

- financial incentive would motivate them even more to work remotely.
- The majority of respondents could not accurately evaluate the following statements (a rating of 3 was chosen - neither agree nor agree): that they miss personal contact with employees and management staff; that when working remotely, the boundaries between professional and private life are blurred; that they are overloaded with information and communication when working remotely; that when working remotely, they face challenges in how to motivate themselves to work; that when working remotely, they face work disruptions from other people in the household; that they are more productive in a classic workplace and that they do more in classic work. Also, the majority did not accurately evaluate the claims that, based on better conditions or additional benefits, they would rather continue working in the office than from home.
- The majority of respondents did not agree (rating 2) with the following statements: that remote work leads to a lack of trust on the part of management staff; that there is a lack of team spirit when working remotely; that they have more communication problems when working remotely; that there is a lack of response from superiors when working remotely; that, when working remotely, they doubt that their finished work is properly assessed; that there are limitations to career possibilities when working remotely; that when working remotely, there is a lack of important information for work and important information regarding the company itself, and that when working remotely, there is a lack of inspiration and challenges at work.

We entered the obtained results into the model and formed the following conclusions with the help of decision trees:

- that digital literacy plays a key role in the very nature of the work of an employee in the IT sector. The obtained results tell us that it is basically a criterion that must be guaranteed so that an employee in such a position can perform his job well;
- that when working remotely, employees can usually have a great deal of control over the performance of tasks or over their schedule. This flexibility can have a positive effect on work itself, as well as on productivity and satisfaction. Namely, their working hours can be adjusted in such a way that they better balance their professional life with their private life;
- that employees are aware of saving time and reducing certain work-related costs, and that these two factors can have a significant impact on employee productivity;
- that employees attribute reduced distractions at work (e.g., conversations and social interaction) as an opportunity for greater productivity;
- that the respondents believe that they are proven to be more productive when working remotely, which is also influenced by self-motivation and job satisfaction;
- that employees are aware of the challenges posed by working remotely, but are not entirely convinced that the additional benefits at the workplace would convince them to return to the classic way of working;
- that employees are aware that their productivity also depends on their capacity and their ability to work remotely and
- that employees have no problems with the lack of contact between employees and management.

The measuring instrument in the continuation of the qualitative research was an unstructured interview. Figure 2 shows the design of our research using a focus group.

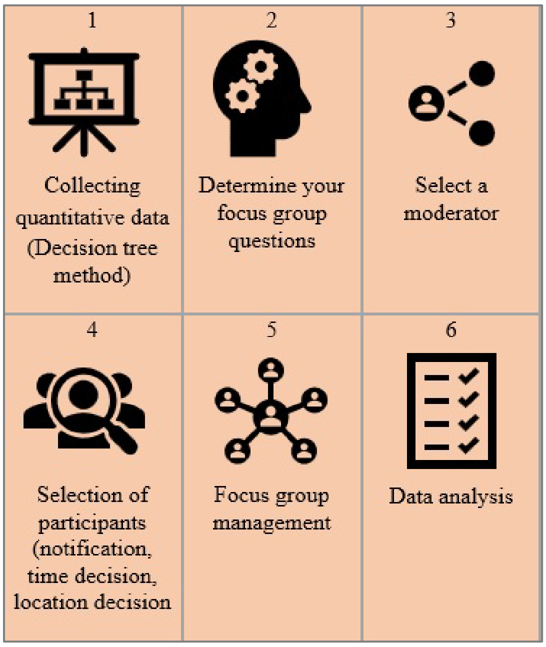

1
Collecting quantitative data (Decision tree method)
2
Determine your focus group questions
3
Select a moderator
4
Selection of participants (notification, time decision, location decision
5
Focus group management
6
Data analysis

Figure 2: Preparing a design for a focus group

The questions were based on the results obtained using the decision tree method. A focus group with seven ICT company employees chose the MS Team application for a group interview that lasted 1 hour and 10 minutes. The contractor asked pre-formulated questions and sub-questions and encouraged all participants to participate, especially those who were more reluctant, so that everyone in the focus group expressed their opinions and that everyone's opinions were equally represented. Data collection was carried out with the fundamental ethical principles of qualitative research. All participants approached the interview voluntarily, with the possibility of discontinuing participation at any stage of the research. Figure 3 shows the data collection process using a focus group.

Question: **"How important is the level of complexity of digital literacy when working remotely in the IT sector"?** The percentage of participants who considered the level of complexity of digital literacy extremely important is: 71.4%, and the percentage of participants who considered it a rather important factor is: 28.6%. The analysis reveals that the majority of participants (71.4%) believe that the level of complexity of digital literacy is extremely important when working remotely in the IT sector. The rest of the participants (28.6%) consider it quite important. When answering this question, the participants agreed that it is difficult to talk about the importance of digital literacy in their work, which actually

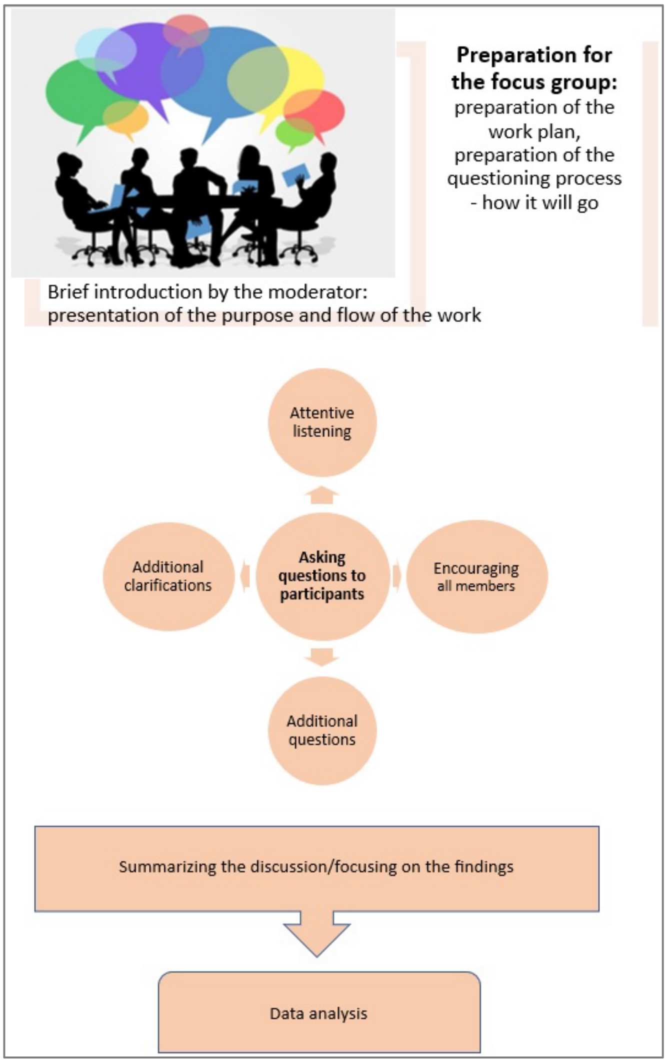

Preparation for the focus group:
preparation of the work plan, preparation of the questioning process - how it will go
Brief introduction by the moderator: presentation of the purpose and flow of the work
Attentive listening
Additional clarifications
Asking questions to participants
Encouraging all members
Additional questions
Summarizing the discussion/focusing on the findings
Data analysis

Figure 3: Focus group process and data collection

comes from their ability to work with computers and programs. They added that it is also important that they know how to manage information, which includes understanding, using and transferring it.

Question: **"Could telecommuting productivity be higher simply because employees can set their own time and thus actually spend less time being unproductive"?** The percentage of participants who consider it important to be able to choose their time when working remotely amounts to is: 60%, and the percentage of participants who consider that they are more productive due to adjusting their time and saving time is: 40%. The analysis reveals that the majority of participants (60%) believe that it is important to be able to choose the time when working remotely, that is, to be able to determine when they will carry out some work and that as a result, they are more productive. The rest of the participants (40%) believe that they are more productive due to adjusting their time and saving time that does not have to be used for small things (e.g., commuting, short speeches at the workplace, etc.).

Question: **"Could telecommuting, with its potential benefit of improving work-life balance, contribute to the new normality of doing and accepting telecommuting for jobs not only in IT, but also in other sectors"?** The percentage of participants who think that remote work in the IT sector has other advantages is: 50%, and the percentage of participants who think that in the case of other sectors this might be an advantage they need is: 50%. The analysis reveals that half (50%) of the participants believe that IT has other more important advantages when working remotely, such as saving time, working at home, working in a quiet environment, etc. The remaining participants (50%) believe that remote work has already become normalized in the IT sector and that the stated advantage, i.e., greater balance between work and private life, may be the advantage that will lead to normalization in other sectors as well. All the participants listed the industries in which they believe remote work would also make sense, at least as an additional option: work in the field of finance and banking, medicine and pharmacy, law, the field of art.

Question: **"Does the reduction of mobility have certain disadvantages, as in driving to work having some intrinsic value?"** The percentage of participants who do not miss driving is: 57,1%, and the percentage of participants who miss some aspects of driving is: 42.9%. The analysis reveals that the majority of participants (57.1%) do not miss driving to work, as they can save time and are not stressed, e.g., due to traffic rush hours and the like, which means that the reduction in mobility has no negative consequences for them. The remaining participants (42.9%) think that they miss certain aspects of driving, such as e.g., independence, isolation in one's own little world, etc.

Question: **"How should an employer in the IT sector maintain greater productivity of employees when working remotely?"** The percentage of participants who consider maintaining connections between employees, regular software updates and relaxed communication to be important is: 100%; the percentage of participants who suggest financial assistance when setting up a home office is: 28.6%; the percentage of participants who mention praise for a job well done as a productivity factor is: 57.1%. The analysis shows that all or almost all participants agree on the importance of maintaining employee connectivity, regular software updates and relaxed communication to maintain productivity when working remotely in the IT sector. In addition, (28.6%) of participants suggest providing financial assistance in setting up a home office, while (57.1%) mention the importance of praise for a job well done. These findings highlight key ideas and perspectives shared by focus group participants.

Question: **"Could employees have less problems with lack of team spirit due to the very nature of the work, where employees with a certain level of personality dominate"?** The percentage of participants who agree that they do not feel a lack of contact when working remotely and that there is no reduced team spirit is: 100%, the percentage of participants who believe that their relationship level is at a high level, as technology allows them to have regular contact, which is just as good, even if they don't have that much physical contact: 100%, and the percentage of participants who declare that they often communicate when working remotely and have normal, friendly and relaxed conversations: 100%. The analysis of the participants' answers showed that all participants, regardless of the nature of their work, believe that working remotely does not affect the lack of team spirit. Also, everyone is of the same opinion that technology makes it possible to maintain a high level of relationships and communication that supports team activities.

Question: **“Which characteristics of remote work could be highlighted as demonstrable advantages for an employee in the IT sector”?** The percentage of participants who highlighted the flexibility of working hours as an advantage of remote work is: 100%, the percentage of participants who highlighted familiarity and greater comfort as an advantage is: 60%; the percentage of participants who highlighted geographical freedom as an advantage is: 50%; the percentage of participants who highlighted fewer distractions and more silence as an advantage is: 50%; the percentage of participants who identified fewer distractions and more silence as an advantage is: 40%; the percentage of participants who highlighted reduced stress from driving as an advantage is: 40%; the percentage of participants who highlighted the ability to achieve a greater balance between professional and private life as an advantage is: 20%. The analysis of the responses of the participants showed that the flexibility of working hours is universally recognized as the most important advantage of telecommuting for employees in the IT sector. In addition, familiarity, comfort, geographical freedom, fewer distractions, greater quietness, lower costs and reduced driving stress are also important benefits mentioned by more than half of the participants.

Question: **“How could employers in the IT sector, in the case of the transition of employees to full-time work from home, ensure or be sure that the employees will cooperate enough, that they will be creative enough themselves, or able to do this work for a long time without negative consequences for productivity?”** The percentage of participants who agree that effective communication is key to ensuring cooperation, creativity and productivity of remote employees is: 100%; the percentage of participants who consider it necessary to monitor work, which can be introduced by the employer with a time calendar and recording of work progress, amounts to: 80%; the percentage of participants who believe that the weekly work report is important for good cooperation, creativity and motivation of employees is: 80%; the percentage of participants who believe that live meetings are key to promoting creativity, good work and preventing loneliness and negative effects on mental health is: 60%. The analysis of the participants' responses showed that effective communication is crucial for employers in the IT sector in ensuring the cooperation, creativity, and productivity of remote employees. Work control, weekly reports and live meetings are also important elements to maintain a high level of productivity and creativity and prevent possible negative influences such as loneliness.

Question: **“How should employees in the IT sector motivate themselves when working remotely?" and "How should an employer motivate employees to increase or maintain productivity when working remotely?”.** Percentage of participants who agree that the motivation of employees in the IT sector partly related to the activity of the work itself, which they like to perform, as the work is a kind of hobby for them, amounts to: 100%; the percentage of participants who believe that the employer could additionally motivate them with an allowance for exceeding the work norm (faster performance of work, flexibility of working hours and customization are the factors that enable exceeding the norm compared to working in an office according to the participants); the percentage of participants who believe that the employer could also motivate them through depreciation of the equipment (by providing the necessary equipment by the employer and reimbursing the costs for wear and tear of the equipment) is: 100%. The analysis of the participants' answers showed that the motivation of employees in the IT sector to work remotely is partly related to the nature of the work they do and it is a kind of hobby for them. In addition, the employer could only motivate them additionally with an allowance for exceeding the work norm and by providing and amortizing the necessary equipment. The participants pointed out only two concrete factors of motivation, which are essential for them, because they are primarily motivated by the fact that they like doing their work.

Question: **“What are the most important challenges faced by employees in the IT sector?”** Percentage of participants who agree that employees in the IT sector face challenges arising from the influence of external factors (they believe that working remotely is more difficult to manage and solve problems and continue working when things go wrong) is: 100%, the percentage of participants who believe that external factors in remote work pose challenges, such as equipment failures, power outages, network outages, etc. (they believe, that these problems are slower or more difficult to resolve at home than in the office) is: 100%; the percentage of participants who agree that taking care of the security of equipment and programs is a challenge when working remotely (they state that employees must pay attention to data protection, their intervention and safety of domestic technology equipment) is: 100%; percentage of participants who find writing reports a challenge (they state that writing meaningful reports that the employer understands is sometimes difficult, as some tasks are self-evident or difficult to explain) amounts to: 57.1%. The analysis of the responses of the participants showed that employees in the IT sector face challenges when working remotely, arising from the influence of external factors. Problems such as equipment failures, power, and network outages present obstacles to smooth operation. In addition, taking care of the security of equipment and programs is also an important challenge. Writing reports is also a challenge, as it is necessary to describe the work in a clear way, despite self-evident or difficult-to-explain tasks.

3 Discussion

The attempt to work from home, which was actually not a real test, as employers and employees were forced to make such a decision due to measures during the COVID-19 pandemic, brought many positive effects. Employers and employees alike have gained valuable insights. Some research has even shown that telecommuting employees were so satisfied that more than half of them would rather find another job than continue working exclusively in offices. More than half of the employees based on these still want to work remotely, although the hybrid model is also very attractive to employees. Still they are less interested in the existing way of work, therefore a return to the classic way of working in the office is not expected (Sokolic) [9].

In the future, we can therefore expect more remote work, according to promising studies of working from home, however few employees (in sectors where this is possible) will work exclusively in a traditional workplace. Even the hybrid model, which also existed before the pandemic, promises many possibilities, the classic way of working certainly cannot be completely replaced by remote work only. In the long term, along with the advantages of remote work, the negative factors can also become increasingly apparent. Among these are certainly the reduction of social contacts and the emergence of the need for interaction.

The question remains what impact any negative factors will have on the ICT field. At least in the current study, we did not find that any of the negative factors would cause any problems for the employees. The employees had challenges, but they did not agree with the claims about the lack of mutual contacts. They also do not agree with other negative effects, such as a reduction in the connection between them and the management staff. Since in our case it is a sector where many people who have introverted personalities work, it would be very interesting to investigate how remote work would be reflected over a longer period.

In the case of our study, we simultaneously started from theoretical findings, other recent research, and modern approaches, as well as from the results of our research, of which the presented hypotheses were established.

With H1: *The perception of remote work depends on the field of work*, we started from the assumption that in the case of IT companies, it is easier for employees to work from home due to the very nature of the work, which already basically attracts a certain type of personality. According to the results obtained from our research, employees in IT companies want to work remotely because they want more flexibility in their working hours. It is also a workplace where employees are increasingly aware that they do not need to move to a well-paid workplace, because there is no longer any geographical limitation when working remotely. Employees in this position also believe that they are more productive when working remotely and accept their remote work very well. In their work, they did not perceive a lack of personal relationships, problems in communication or a lack of information. According to other studies, where other jobs or areas of work are analyzed, the research also shows good results and maintaining or even increasing productivity, but in certain cases outstanding factors can still be traced. Some employees faced a lack of social connection, which in turn led to stress, problems with psychological well-being, and even depression for some. **From this we can conclude that the experience of remote work depends on the individuality of the employees, their preferences, personality type and perhaps not so much on the field of work itself.**

With H2: *The positive effects of telecommuting in the case of the IT sector can affect productivity, and in the case of a hybrid model these could increase even more*, we started from the assumption that both approaches have their advantages and that combining all advantages can lead to even better results. Despite the fact that employees are convinced that they are more productive at home, problems can arise with the management itself. Sometimes productivity is not so easy to measure, and this can lead to a lack of transparency and, as a result, less trust from management towards employees. It is true, however, that a highly productive employee is most likely equally effective both in the office and at home. In our research, we found quite a few advantages that employees point out when working remotely, among which the flexibility of working hours stands out in particular. However, the respondents were not so sure that they would much rather work only remotely than in a classic workplace, when we presented them with the possibility of additional benefits in a classic workplace. If remote work offers flexibility of working time and place and other advantages, on the other hand, it also brings challenges, such as facing and solving problems and lack of social interaction. In this case, the hybrid model seems like a solution that can offer the best of both worlds. Greater flexibility when working remotely, a better balance between professional and private life, as a result, can affect higher productivity. Working in an office also has its advantages, it is a more suitable place for collaboration, creating good relationships, etc. In any case, the period during the pandemic made it possible for employees to experience remote work and compare it with the traditional model. More freedom, greater flexibility of work, there is no going back, many are even ready to change their workplace if it only required them to do classic office work. However, it is still only an experience at a certain time and if we look at it more broadly, face-to-face personal communication cannot be recreated virtually, even work relationships and culture are stronger when they are built physically. **Therefore, the hybrid model is something that can have a long life and something that may really increase the positive effects on productivity.**

With H3: *Employees see many advantages in remote work, but the standard form of work is also close to them, which means that the hybrid model could be the most effective form of work in the IT sector*, we started from the assumption that employees in remote work otherwise see many advantages, but they are not entirely sure that they no longer want to work in the classic way as well. The need for freedom, adaptability, geographical restrictedness, but also for personal communication and socializing is something that puts the hybrid model as a solution for companies that want to adapt to the post-pandemic world. It is a model that offers advantages that include employee satisfaction because they can adjust how they will do their work, but at the same time it ensures that there is no lack of personal contacts and successfully discourages employees from possible isolation and deterioration of mental health. The flexibility of hybrid work allows employees what they have generally wanted for a long time, greater balance, and coordination between their professional and private lives. Despite the fact that the hybrid model appears to be the most efficient form of work, it should be noted that this model will also require careful preparation and planning, and companies will have to adapt to uncertainties in the future as well. Of course, it also depends on the individual employee, for some it will suit them to work from home for part of the week and in the office for the rest, while some may want to work entirely remotely.

4 Conclusion

COVID-19 contributed to the fact that the trend of remote work took hold. Remote work has been known for a long time, but companies did not test it to such an extent until they were actually forced to introduce this form of work in the last two years due to the pandemic. In the case of ICT companies, it is work that is already done from home in many parts of the world, but this time the companies and employees were not ready for an immediate change. The need for quick action and changing conditions did not allow for much planning and preparation. We can certainly talk about the fact that no one, neither the employer nor the employee, was completely ready for the transition. The change of the workplace, processes, resources and the preparation of people was fast, but according to the analysis carried out, the employees responded very well in the case of ICT companies. Employees (also employers) gained experience and checked how they can coordinate their wishes, capabilities and challenges when working remotely. The conducted research showed primarily positive effects on productivity and employee satisfaction with this type of work. The trend of working from home also shows prospects for the future, although a new perspective must be taken into account. The employees were satisfied during the period of working from home and see many advantages in this type of work. However, it was in fact, a period when we could not perform our work in the classic way due to the requirements and measures. Here we can also refer about the satisfaction of being able to do their work without any particular fear of being out of work because it cannot be done or is done under strictly defined conditions. Remote work during this period also offered some security due to the familiarity of the environment itself, so that serious social contacts did not even occur, at least not in the area where we performed the analysis. The question is how it will be in the future. Despite the results, which showed that employees did not miss personal contacts with colleagues and management, it is necessary to consider for the future, e.g., about the need for social contacts, means to maintain this type of work, etc., regardless of the fact that research has also shown that many introverted personalities work in this field. Remote work could continue, but according to the latest research, it appears that the most advantageous way at the moment would be a hybrid model. It is a model that simultaneously offers employees more flexibility while maintaining a certain level of control and stability for the employer. This way of working is certainly more acceptable also because restrictions disappear and people no longer fear and desire to be closed off due to health conditions. The hybrid model allows employers to experiment with solutions and find out what is best for employees and for the company/organization.

Acknowledgment

The author would like to express gratitude for the invaluable support and guidance provided by Matej Mertik and Mladen Radujković, mentors and co-mentors of the research work that have greatly contributed to the success of this research within the PhD study program in Project Management at Alma Mater Euroaea ECM.

References

[1] Tri Wulida Afrianty and I. GustiLanangSuta Artatanaya, and John Burgess, "Working from Home Effectiveness During Covid-19: Evidence from University Staff in Indonesia," *Asia Pacific Management Review*, 27(1):50–57, 2022.

[2] William Arruda, "6 Ways COVID-19 Will Change the Workplace Forever," https://www.forbes.com/sites/williamarruda/2020/05/07/6-ways-covid-19-will-change-the-workplace-forever/?sh=5d7be8af323e, 2020.

[3] Jose Maria Barrero, and Nicholas Bloom, and Steven J. Davis, "*Why Working from Home Will Stick*," NBER Working Paper Series, No. w28731. Cambridge, Mass: National Bureau of Economic Research, 2021.

[4] John Cook, Yishay Mor, and Patricia Santos, "Three Cases of Hybridity in Learning Spaces: Towards a Design for a Zone of Possibility," *British Journal of Educational Technology* 51(4):1155-1167, 2020.

[5] Kantha Dayaram and John Burgess, "Regulatory Challenges Facing Remote Working in Australia," V *Handbook of Research on Remote Work and Worker Well-Being in the Post-COVID-19 Era*. Pennsylvania: IGI Global, pp. 202, 2021.

[6] Kanwar Muhammad Javed Iqbal, Sergey Evgenevich Barykin, and Farooq Khalid, "Hybrid Workplace: The Future of Work," V *Handbook of Research on Future Opportunities for Technology Management Education*, Pennsylvania: IGI Global, pp. 28-48, 2021.

[7] Sean Peek, "Communication Technology and Inclusion Will Shape the Future of Remote Work," https://www.businessnewsdaily.com/8156-future-of-remote-work.html, 2022.

[8] Kyle Shaner, "The Future of Work: What's the Future of Work from Home?" https://www.uc.edu/news/articles/2022/09/the-future-of-work--whats-the-future-of-work-from-home.html, 2022.

[9] Danijela Sokolic, "Remote Work and Hybrid Work Organizations," V *78th International Scientific Conference on Economic and Social Development*, Portugal, pp 202-213, 2022.

Tomaž Kokot is a PhD student in Project management at Alma Mater Europaea - ECM and manager in the field of ICT with valuable experiences in public services with a position as general director of Post Slovenia. His research focuses on management, human resources and use of new and digital services in fields.

Predictive Water Quality Modelling Using ARIMA and Water Parameter Forecasting Model (WPFM) for Godavari River, Maharashtra, India

Sucheta Sable/Kakde* and Rajesh Kherde†
D.Y. Patil University, Ambi, Pune, INDIA

Abstract

The Godavari River in Maharashtra, India, is used as an example in this paper to demonstrate the recital use of machine learning methods, including auto regression with mean average and random forest regression. Water data are gathered for modelling from the Hydrological Data Users Group in Nasik, Maharashtra. The procedure used in this article demonstrated tension during the creation of Python code and its transfer to the operation data to predict output. Comparing the two models, ARIMA is used to forecast the places' subsequent six progressive values. After comparing the models' outputs, the RSME of Water Parameter Forecasting Model had a score of 0.90, while the ARIMA received less than 0.5, it was declared to be a reliable model for forecasting the following six values. In order to compare the predicted and generated results, samples from the relevant study sites are gathered and tested in the lab concurrently with the machine learning process.

Key Words: ARIMA, water parameter forecasting model, Jupyter notebook, Godavari river.

1 Introduction

The majority of nations recognize surface water quality as a delicate and important problem. The effect of surface and groundwater quality on human health, aquatic life concerns, and related factors is significant. [12] Water has a major impact on life sustainability which is determined as a key element in our environment. Water found in sea and land plays an important part in day to day life activities such as drinking, agriculture, industrial and other uses [14]. Ground and surface water quality is declining as a result of human-made activities like industrial refuse, agri-/aquaculture, and discharges from other uses [2]. Hydro chemical property analysis is a key component in identifying the quality of water for domestic, commercial, and irrigation uses. Remote sensing, GIS, and statistical analytical tools were used to identify the variables. It regulates the flow system at various Wheeler Lake Basin sites in Northern Alabama. The research evaluated ground and surface water after human consumption by using the water quality index (WQI) method. In various locations around the globe, numerous investigations are carried out to address quality standards and issues. With the aid of engineering studies on rivers, a number of investigations into earth science is the evaluation of water quality parameters. Water found in sea was used to evaluate the status of concerned locations. The most frequent occurrence when it comes to earth science is the evaluation of water quality parameters, water quality, the transport of sediment, and the transmission of pollutants. Human socio-economic growth depends upon readiness of quality water. [11] A rapid increasing in population and thus expansion in agriculture and industries has shown us how that quality water is becoming difficult to achieve. [5] Additionally, measuring is divided into two categories: components of water quality and the spread of pollutants and their mechanisms. Physicochemical analysis was used to classify the water quality parameters as a result of environmental engineering breakthroughs. BOD, COD, pH, temperature, K, Mg, Na, TDS, and other assays are among them [1]. Departments of the federal, state, and local governments as well as the Hydrological Data Users Group are mandated to routinely assess the water quality parameters.

Additionally, the station points will contain fundamental data for creating conservation initiatives. The availability of water quality parameters made it possible to review with the aid of an earlier study [13]. The benefit of using soft computing methods and having access to time series analysis is that the majority of engineering researchers today have conducted the necessary research to make accurate predictions of future quality parameters using mathematical formulas [6]. The MLP model has accurate findings in predicting the parameters when using adaptive neuro-fuzzy inference system (ANFIS), radial basis network (RBF), and multilayer perception (MLP) methods to forecast the water quality parameters [3]. It was suggested to use a model to plan the spread of classified surface water in accordance with the quality of Iran's locations using probabilistic support vector machines (PSVMs) and GIS techniques [10]. Numerous case studies were used to predict the water quality parameters in a different research [4]. The prediction of quality factors with internal relationships used time series analysis and statistical models. Numerous research papers make use of the quality evaluation of water and predictive analysis for the planning of projects related to water conservation. The Godavari River in Maharashtra, which has been identified as the main contributor to the area's surrounding canals, was the subject of the literature reviews used in the development of the current paper's water quality model. Both the ARIMA and the Water Parameter Forecasting Model are used for parameter prediction within the chosen region. Additionally, Jupyter notebook and scripts are created and included to manage the study's data.

* PhD Research Scholar, Department of Civil Engineering.
† Professor, Department of Civil Engineering.

1.1 Proposed Methods and Materials

This paper uses machine learning methods to suggest the distinctive structured code in the Jupyter notebook. There are two program models included: ARIMA and Water Parameter Forecasting Model. The introduction to the study area is followed by a presentation of the periodic water quality statistics and their ranges. The overview of the used time series models is then presented, and the obtained findings are contrasted.

2 Study Area and Data Collection

The Godavari waterway, which has its source in Trimbakeswar, Nasik, Maharashtra, India, is the second-largest waterway in that country. It flows through the regions of Andhra Pradesh, Madhya Pradesh, Karnataka, and Orissa. The waterway, which flows through Nasik City, is 82% domestically and 18% industrially polluted. The study spans a distance of 350km along the river, beginning in Kushawart Trimbakeswar and ending in Saikheda Village, where the river joins the city. The river's water was sampled from ten different places, and the samples were then examined in a lab run by the Hydrological Data Users Group for indicators of water quality.

The information on water quality is gathered from the Hydrological Data Users Group, Nasik (HDUG) database in India, which is maintained by the authorities who are in charge of overseeing the various aspects of water quality. From 2011 to 2021, a total of 10 years are used. The statistics for the monthly and yearly averages were only available in a few places. The four chosen sites are shown in Figure 1 along with the parameters for the water quality. Regarding the Hydrological Data Users Group, Table 1 presents an overview of the elements that make up water quality. BOD, COD, pH, temperature, K, Mg, Na, TDS, and DO are the parameters this research takes into account.

3 Predictive Modelling

Making predictions based on historical data can undoubtedly assist in resource management and raise the standard of the water supplied to the community.

The dependent variable's future value can be predicted using a unique set of techniques and methods that are accessible in the machine learning domain [15]. The ideas of the ARIMA model, which capture the essence of time series analysis, are used in this paper. The back end date given must be checked with the various types of components, such as trend, noise, and seasonality, in order to run the ARIMA model. The time periods are listed and depicted in Figure 2. The data must be stationary according to the statistical model for a time series collection. The following list includes the codes that are necessary to complete the time series analysis and statistical model.

A specified "excel" file containing the data has been imported into the Jupyter notebook, and the command window appears as shown in Figure 2. Making a good forecast model requires making sure the time series data is stationary.

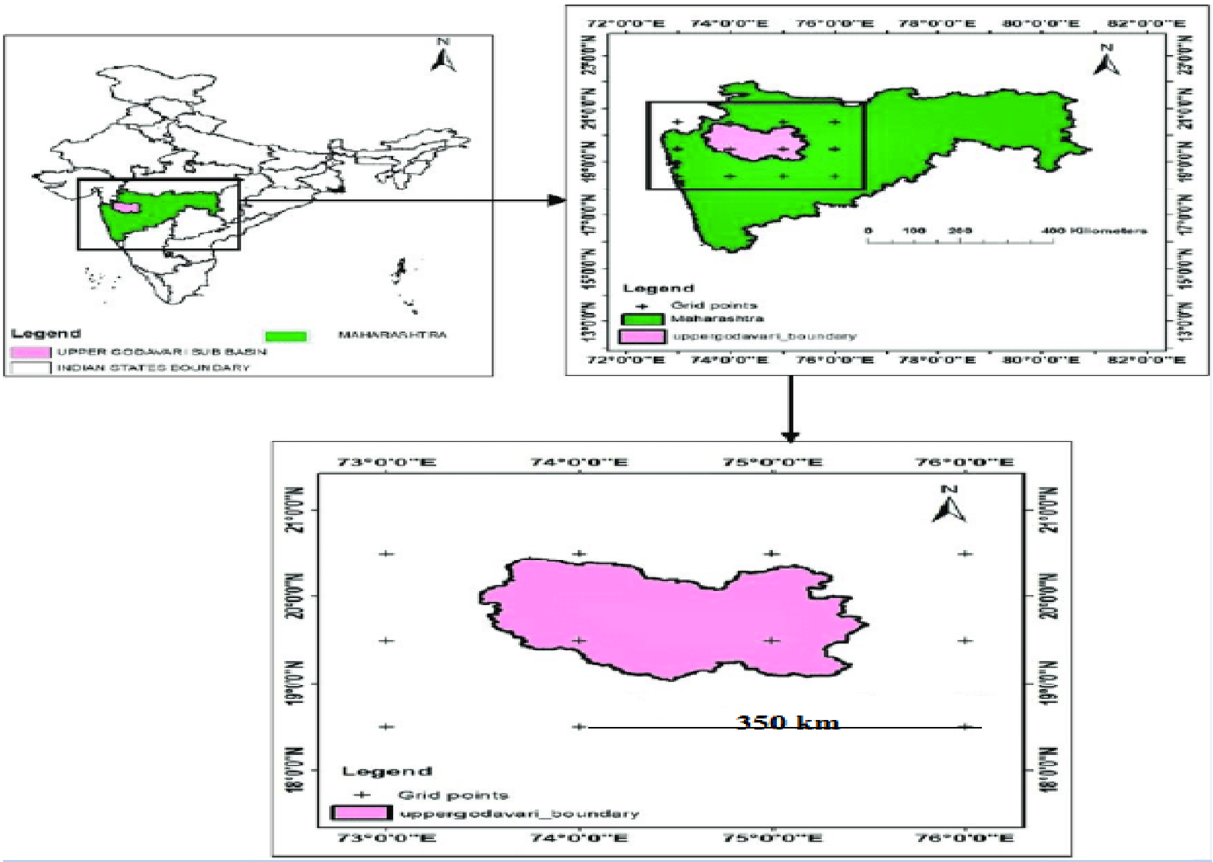


Figure 1: Location of the study area

Table 1: The summary of water quality parameters by HDUG at Godavari River

Annual Average Values	BOD	COD	DO	Ph	Temp	SS	K	TDS
Date	Average							
2012	18.9	24.4	1.8	7.28	25.7	0.65	2.774286	344.6
2013	21.1	21.6	2.5	7.24	25.7	0.59	2.685714	333.6
2014	17.4	23.4	3.3	7.25	25.7	0.54	2.315714	337.9
2015	16	25.4	3.3	7.24	25.8	0.5	2.128571	340.6
2016	17.6	28.1	3.9	7.23	25.8	0.47	1.957143	345.6
2017	19.1	29.3	4.3	7.25	25.7	0.51	1.471429	340.3
2018	19.8	30.3	3.8	7.26	25.7	0.49	1.071429	319.3
2019	20	29.9	3.7	7.27	25.6	0.48	1.342857	323.2

The given time sequence’s future values, after being evaluated by its past and current statistical values, can be predicted using the ARIMA model. The ARIMA model can be divided into three categories, with “AR” standing for autoregressive, which takes prior values into account when a variable is evolving. While “I” reverses the integration part of data values evolved under the process of differentiation between the values of present and past, “MA” showed the linear combination of errors and values with a moving average part of regression at various times of past values. In the end, the ARIMA model's features can help the data match the range shown in Tables 1 and 2.

3.1 Model Determination

Mushtaq analyzed the data using the rolling statistics and augmented the Dick fuller test (ADF) [9]. Further, non-stationary data, an autocorrelation plot with decay shall be viewed in the window. Analysis of the excel data of different parameters was done to understand the pattern present between data.

The following algorithm is suggested for second-order differencing based on the outcome produced following the first-order differencing of non-stationery data:

3.2 Parameter Estimation and Analysis Using Model

The regression model is created by averaging the parameter data and using the date “x” as an independent variable and the parameter “y” as a dependent variable.

The Autoregressive Integrated Moving Average model, or ARIMA model, is a well-liked time series forecasting technique that incorporates moving average, autoregressive, and differencing components.

The general formula for an ARIMA(p,d,q) model is:

Y_t = c + ϕ_1(Y_t-1 - μ) + ... + ϕ_p(Y_t-p - μ) - θ_1ε_t-1 - ... - θ_qε_t-q + ε_t

Where:

- Y_t is the value of the time series at time t
- c is a constant term (the intercept)
- ϕ_1, ϕ_p are the autoregressive coefficients (AR terms) for lags 1 to p
- Y_t-1, Y_t-p are the lagged values of the time series
- μ is the mean of the time series
- θ_1, ...,θ_q are the moving average coefficients (MA terms) for lags 1 to q
- ε_t-1, ..., ε_t-q are the lagged errors or residuals
- ε_t is the error term or residual at time t
- is the degree of differencing, which is the number of times the series is differenced to achieve stationary.

3.3 ARIMA Determination Step

The Autoregressive Integrated Moving Average (ARIMA) model is a popular time series forecasting technique used to analyse and predict data that exhibits a trend, seasonality, or other patterns. The process of determining the ARIMA model for a given time series involves the following steps:

1. Visualize the data: Plot the time series and examine its behavior. Look for patterns, trends, and seasonality.
2. Stationarity check: A stationary time series has constant mean and variance over time and is necessary for ARIMA modelling. Check for stationarity using statistical tests like Augmented Dickey-Fuller (ADF) or KPSS test. If the series is not stationary, apply transformations like differencing or seasonal differencing.
3. Identify the order of differencing: If the data is not stationary, apply differencing until it becomes stationary. The order of differencing is the number of times the data is differenced. The order of differencing can be identified visually or using the autocorrelation function (ACF) and partial autocorrelation function (PACF) plots.
4. Determine the order of Autoregressive (AR) term: The AR term is the number of lagged values of the time series that are used to predict the current value. The order of the AR term can be identified using the PACF plot.
5. Determine the order of Moving Average (MA) term: The MA term is the number of lagged forecast errors used to predict the current value. The order of the MA term can be identified using the ACF plot.
6. Model selection: Combine the identified orders of differencing, AR, and MA terms to form a candidate ARIMA model. Select the best model based on statistical measures such as AIC (Akaike Information Criterion), BIC (Bayesian Information Criterion), or

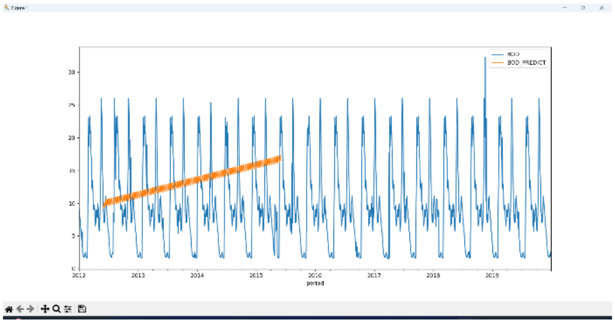
MSE (Mean Squared Error) on a validation set [7].

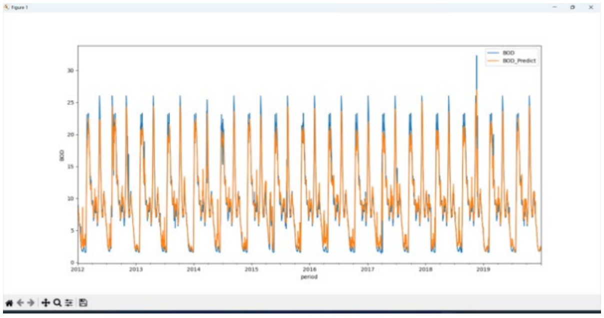
7. Model validation: Validate the selected model using statistical tests and plots, such as residual analysis, Q-Q plots, and Ljung-Box tests.

8. Forecasting: Use the selected and validated ARIMA model to make predictions on future values of the time series.

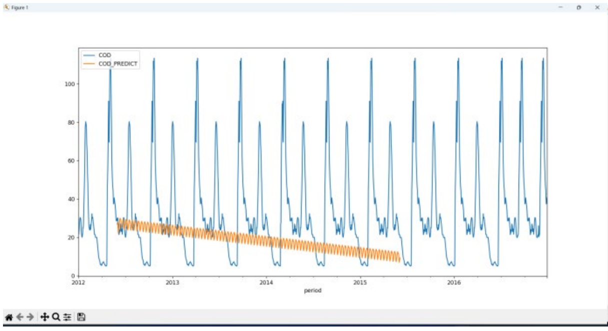
ARIMA **WPFM**

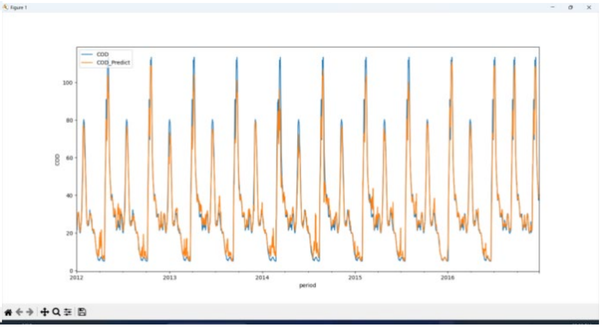
BOD



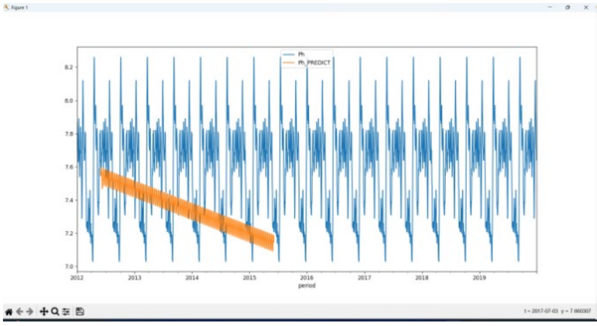


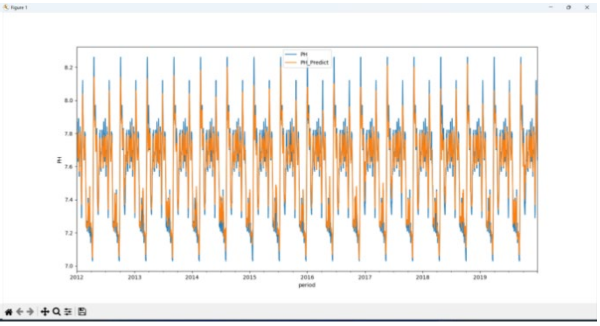
COD



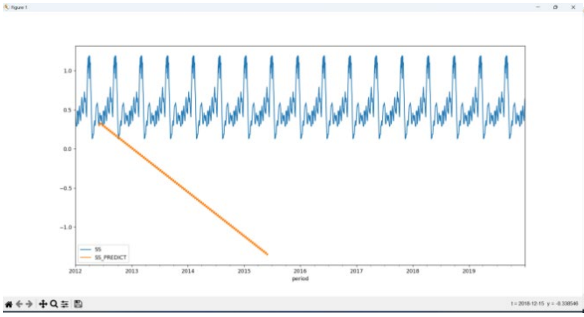


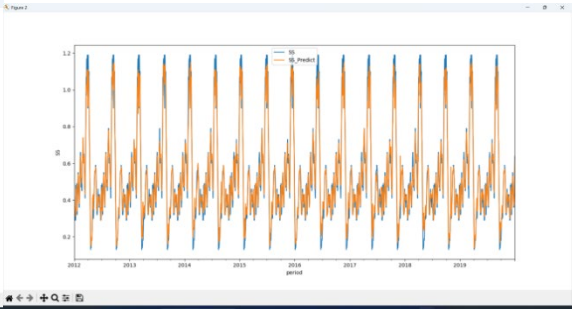
Ph

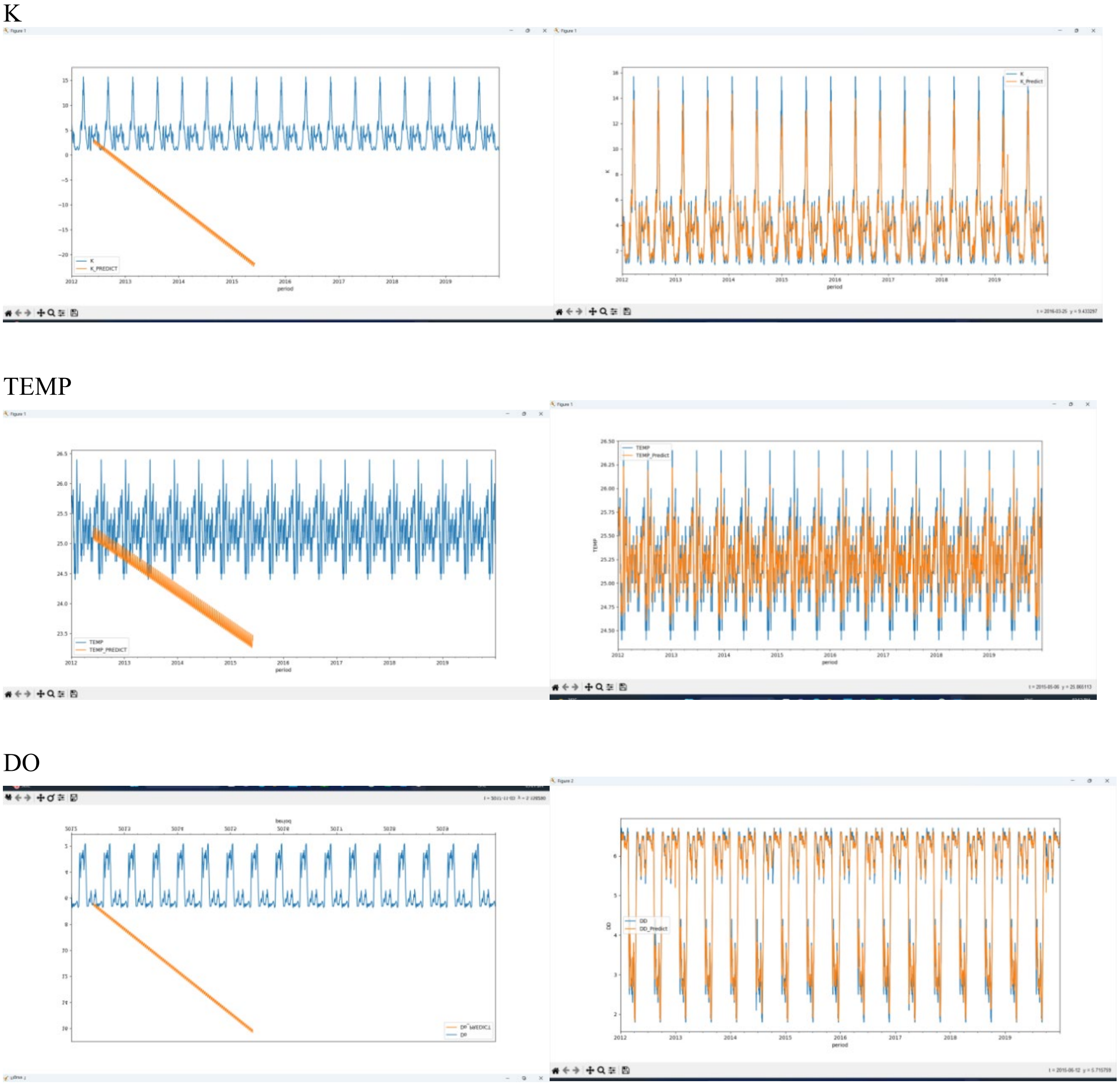




SS






K

TEMP

DO

Figure 2: Predictive modelling using water parameter forecasting model

3.4 Water Parameter Forecasting Model (WPFM)

WPF model is used as a random regression method, which is based on the generate tree structure to estimate the parameter prediction. The process of random forest regression entails building several decision trees, which are then combined to produce a model that is more reliable and precise. A random subset of the accessible data and features is chosen to create each decision tree [8]. The random pick contributes to improving the model's generalizability and decreasing over fitting. The regression model is used in statistical analysis to forecast the parameters "*y*". It uses past time series data to forecast the output. Random forest regression model it is an ensemble technique to take the sum of the all the estimated output.

3.5 Model Determination

$$X_{newt} = X_t + X_{t+7} \quad (1)$$

To find the output by Random Forest Regression

$$\text{Standard Deviation} = \sum (X - X_{MEAN})/N \quad (2)$$

Find the probability of each attribute

Probability Distribution = $\sum$ P(X) × Standard

Deviation(X) (3)

3.6 Main Library:

From Sklearn.ensemble.Random Forest.
Pandas
numpy
Sklearn.model-selection
matplotlib.

3.7 Regression Tree Generation using Machine Learning (ML) Involves the Following Steps

1. Data preparation: Collect and pre-process the data, which involves cleaning, transforming, and normalizing

the data. Split the data into training and testing sets.

2. Feature selection: Select the most important features that are relevant to the prediction task. Feature selection can be done using statistical tests, domain knowledge, or ML algorithms.
3. Tree building: The tree building process starts with the selection of the root node. The root node is selected based on the feature that maximizes the difference between the target variable and the feature values. The data is then partitioned into two subsets based on the value of the selected feature. This process is recursively applied to each subset until a stopping criterion is met.
4. Stopping criterion: The stopping criterion determines when to stop partitioning the data into subsets. The stopping criterion can be based on the maximum depth of the tree, the minimum number of samples in a leaf node, or the minimum reduction in the variance of the target variable.
5. Tree pruning: Tree pruning is a process of removing unnecessary branches from the tree to prevent over fitting. Over fitting occurs when the model is too complex and captures noise in the training data.
6. Model validation: Validate the model on the testing set using statistical measures such as Mean Squared Error (MSE), R-squared, or Root Mean Squared Error (RMSE).
7. Hyperparameter tuning: Hyperparameters are parameters that are not learned from the data but are set before the training process. Examples of hyperparameters include the maximum depth of the tree, the minimum number of samples in a leaf node, and the learning rate. Hyperparameter tuning involves selecting the best hyperparameters that optimize the performance of the model.
8. Prediction: Once the model is trained and validated, use it to make predictions on new data.

3.8 Model Development

The first stage in putting machine learning models into practice is the preparation of the data set. The obtained data collection should now be divided into two categories: training the input data and testing. The validation and evaluation of models that have been applied use training and trial data sets, respectively. The process of allocating a subgroup to each category differs depending on the parameter values for time series modelling. In time series, it is best to assume that the history of data gathering has been modelled, and shuffling the data set is not appropriate, though it is acceptable when a feature is present.

Usually in both situations, approximately 75-85% of the data are set aside for confirmation and the resulting 20-30% for confirmation. The implementation of machine learning models, including the tested and best-chosen ARIMA and Water Parameter Forecasting Model, is the following step. The network's capacity needs to be adjusted in the following phase to improve the consistency of the Predictive Water Quality Modelling Using ARIMA and Water Parameter Forecasting Model for Locations of Godavari River. In order to accomplish this, more neurons or secret layers will be added. This method's final two stages also cover the creation of the Initiate software architecture.

4 Results and Discussion

The ARIMA and Water Parameter Forecasting Model are both described in this paper along with their use in predicting the outcomes of water tests for a few water components. An optimal model created in this paper has experienced analysis, and in the later stages, predictions have been made using various time series analysis methods. Two-time series analysis models were modified from separate models, and new code was generated to fit the ARIMA (autoregressive integrated moving average) and Water Parameter Forecasting Model. These two models' various parameters, including temperature, pH, NA, K, COD, BOD, dissolved oxygen, and TDS, are chosen based on historical data from government organizations. In the same way that the predictive values for the year 2030 were predicted, the total data values from the years 2012 to 2019 have been gathered and evaluated. The time series deflection for the predictive values could be seen for both models, which had both done well.

Following the completion of these operations, a number of observations are made to assume the predicted values can be discovered, and the best model that was predicted is then chosen. Personal interest can be used to determine the number of observations, but more backend data can actually help the algorithm perform better. Given that there will be 100 observations in this specific value, the data is split into two sets: one for training and the other for testing. Based on test data values from the two models, ARIMA and Water Parameter Forecasting Model, the projected values will be obtained after the test data; we obtain the predicted values after the test data. We will then plot the graphs for the test data and the predicted numbers after making our predictions.

In Tables 2 and 3, the Water Parameter Forecasting Model achieved 95% accuracy and the RSME of the predicted data for both models was met at 85% accuracy for the ARIMA-model.

Table 2: ARIMA

Performance Measure	BOD	COD	Do	Ph	Temp	SS	TDS	K
RMSE	3.4	20.4	1.34	2.23	11.3	5.65	12.3	5.22
R2_Score	.49	.5	.56	.55	.44	.25	.55	.5

Table 3: Water parameter forecasting model

Performance Measure	BOD	COD	Do	Ph	Temp	SS	TDS	K
RMSE	0.89	1.85	.5	.65	.33	.45	3.233	2.1
R2_Score	.92	.89	.925	.94	.88	.85	.89	.88

5 Conclusion

The model's outcomes in this article were based on a mathematical approach. Utilizing Jupyuter Note, statistical formulas, the code necessary for parameter analysis is imported from the data library, and both the data and the code are changed in accordance with the time series. In the latter part, the predicted values had also met the R2 values, which range from 0.92 to 0.97. As we look at the graphs, the initial values are shown in blue, and predicted values are coded in orange. The seasonality of data sets is thoroughly examined in detail. Data from the last 8 years' worth of time series are seen to be nonstationary. Thus, it appears that the graph generated by the Python-based Jupyter notebook is also nonstationary.

Water Parameter Forecasting Model forecasts values that are closer to the original values than the ARIMA model, when the two models are compared. Compared to the ARIMA model of RSME values, the RSME & R2 values in the Water Parameter Forecasting Model are more comparable. The values discovered through data analysis are shown in the chart below for both the ARIMA and Water Parameter Forecasting Model.

References

[1] M. B. Assad and R. Kiczales, "Deep Biomedical Image Classification using Diagonal Bilinear Interpolation and Residual Network," *Int. J. Intell. Netw.* 1:148-156, 2020.

[2] Ajay D. Chavhan, M. P. Sharma and Renu Bhargava, "Water Quality Assessment of the Godavari River," *Hydro, Nepal*, 5:31- 35, 2009.

[3] S. Emamgholizadeh, H. Kashi, I. Marofpoor, and E. Zalaghi, "Prediction of Water Quality Parameters of Karoon River (Iran) by Artificial Intelligence-Based Models," *Int. J. Environ. Sci.Techno*l. 11(3):645-656, 2013.

[4] S. Heddam, "Generalized Regression Neural Network Based Approach as a New Tool for Predicting Total Dissolved Gas (TDG) Downstream of Spillways of Dams: A Case Study of Columbia River Basin Dams, *USA, Environ. Process*, 4(1):235-253, 2016.

[5] C. T. Hunsaker and D. A. Levine, "Hierarchical Approaches to the Study of Water Quality in Rivers—Spatial Scale and Terrestrial Processes are Important in Developing Models to Translate Research Results to Management Practices," *Bio Science*, 45(3):193-203, 1995.

[6] N. S. Jaddi and S. Abdullah, "A Cooperative-Competitive Master-Slave Global-Best Harmony Search for ANN Optimization and Water-Quality Prediction, *Appl. Soft Compute*, 51:209-224, 2017.

[7] L. Kilian, "Small-Sample Confidence Intervals for Impulse Response Functions," The Review of Economics and Statistics, *MIT Press*, 80(2):218-230, May 1998.

[8] M. Maleki and S. M. Kashefpour, "Application of Numerical Modelling for Solution of Flow Equations and Estimation of Water Quality Pollutants in Rivers (Case study: Karkheh River)," *Civil Environ. Eng.*, 42.3(68):51-60, 2012.

[9] R. Mushtaq, "Augmented Dickey Fuller Test, [online] papers.ssrn.com. Available at: https:// ssrn.com/abstract=1911068, 2011.

[10] M. R. Nikoo, N. Mahjouri, "Water Quality Zoning using Probabilistic Support Vector Machines and Self-Organizing Maps, *Water Resour. Manag.*, 27(7):2577-2594, 2013.

[11] C. C. Obropta, M. Niazi, and J. S. Kardos, "Application of an Environmental Decision Support System to a Water Quality Trading Program Affected by Surface Water Diversions," *Environmental Management*, 42:946-956, 2008.

[12] Y. Ouyang, "Evaluation of River Water Quality Monitoring Stations by Principal Component Analysis," *Water Res.*, 39(12):2621-2635, 2005.

[13] A. Parsaie, A. H. Haghiabi, M. Saneie, and H. Torabi, "Applications of Soft Computing Techniques for Prediction of Energy Dissipation on Stepped Spillways," *Neural Compute. Appl.*, 29:1393-1409, 2018.

[14] Sucheta Sable/Kakde, Rajesh Kherde, and Gauri Patil, "A Review for Water Quality Modelling for a River Basin, The Seybold Report, 17(9):1410-1420, ISSN No-1533-9211, DOI 10.5281/zenodo.7115974, 2022.

[15] Purushottam R. Sarda and Parag Sadgir "Water Quality Modelling of Godavari River, India using Q2kw Soft Tool," Conference: India Water Week 2015: At: Delhi, India, January 2015.

Sucheta Sable/Kakde completed B.E.in civil engineering and ME in water resources engineering. Currently Pursuing Ph.d in the water resources engineering. I am currently working as an assistant professor in Engineering Collage and I am passionate about my job. Email: suchetasable17@gmail.com. Phone no.-+917338308858

Rajesh Kherde (photo not available) is the Principal at D. Y. Patil School of Engineering & Technology, Ambi, Pune. He has a Ph.D. in Civil Engineering with specialization in Water Recourses Management. Prof. Dr. Rajesh Kherde is an academician with Ph.D. in Civil Engineering from Mumbai University and a Gold Medallist of Gujrat University in a Master's study. He has 23 years of experience in the capacity of Principal, Professor and Head of the department for several years. He has worked as BOS member of civil engineering course in faculty of engineering at Sandip University, Nashik. He has successfully handled several responsibilities like NAAC coordinator, IQAC coordinator, Admission in charge and vice chancellor nominee as an observer for admission process.

Hybrid SMOTE and Bootstrap Sampling for Imbalanced Classification in Elderly Health Condition Dataset

Rattanawadee Panthong*
University of Phayao, Phayao, THAILAND

Abstract

The dataset was applied in the analysis of the real-world problem which found that the data lacked balance and had several classes due to the data nature and the data collection limitation. This made the data gained a high term of possible imbalance. When the imbalanced data is used in learning, it may reduce the efficiency of data classification. Thus, this study aimed to manage the imbalanced data via hybrid data-level technique using SMOTE and bootstrap approaches based on one versus all with different learning. The four different learning methods; deep learning, stacking algorithm, random forest, and gradient boosting tree are applied to improve the accuracy rate of the classification model. The elderly health condition dataset was obtained from the Meaka Heath Promotion Hospital of Phayao in Thailand. The experimental results indicated that the HySM_BT50% method gained the highest correctness value at 90.11% (Sensitivity = 0.8469, Specificity = 0.9749, and G-means = 0.9514) when using random forest algorithm as a classifier.

Key Words: SMOTE; bootstrap; imbalance; multiclass; classification; deep learning; ensemble method; elderly health condition.

1 Introduction

Most problems found in machine learning are the imbalanced data and have a lot of multi-classes from the dataset occurring from the distribution of the sample group or unequal label class, for example, one class has the least proportion compared with other classes resulting in predicting or classifying the minority class with having low efficiency and misclassification [2, 3]. This is because the data classification method will give good efficiency when the data is balanced or close to each other. The imbalanced data occurs in different domains such as the problem of the medical cluster classification (cancer data and no cancer patients), the problems of the risk management and the anomaly detection [6]. It is quite difficult to gain the data in each cluster having an equal number.

In real-world problems, the most prevalent problem found is that the multi-class imbalanced data set used the form of the text or the image data. A multi-class problem is the classification problem where the instances of data classification fall into one or more than two classes. Thus, the problem of data classification is a challenging issue for the management of the imbalanced multi-class data. The multi-class classification becomes the main problem of machine learning due to the data having several label classes leading to processing the data more complicated. Moreover, in classifying the multi-class data, it shows that the model efficiency in classifying data relies on the majority vote and the prediction of the new class data [24]. The most common solution, involving multi-class problems which are likely to be harder to predict, is to transform them into several binary problems. One-Versus-All (OVA) decomposes classes into binary and then learners improve the representation of the minority examples. This method is easy and useful for the work requiring to classify the multi-class which focuses on adjusting the data to be appropriate for an algorithm [15, 24].

In general, two data-level approaches used in classifying the imbalanced data are applied to increase the samplings in minority data by over sampling and to decrease the samplings in majority data by under sampling. The technique, most likely used by many experts, is SMOTE (synthetic minority over sampling technique) which is introduced to manage the imbalanced data [12, 19, 23, 26]. It is a way to increase the data in the minority group resulting in spreading the more balanced data clusters in order to make the classification of the minority group much better and the model classification much more accurate. The under sampling approach is a way of reducing the majority sample group to have the same or near amount of the minority sample group leading to adjusting the dataset balance before applied for the training model.

The survey data of the elderly health condition gained from Maeka Health Promotion Hospital, Phayao Province in the total of 1194 instances is divided into 3 classes; class 0 is a group of 610 social bound instances, class 1 is a group of 529 home bound instances and class 2 is a group of 55 bed bound instances. From the data of the elderly health condition, it shows the imbalanced data normally found that a number of healthy people is higher than unhealthy people. In assessing and planning the elderly health care, the minority data is more interesting than the majority data which means that the total of instances of class 0 and class 1 put together is higher than class 2. This will result in classifying the data incorrectly or putting in a wrong cluster and making the classification of the minority group having low efficiency. The adjustment of the imbalanced

*Department of Information Technology, School of Information and Communication Technology, , rattanawadee.pa@up.ac.th, Tel.: +66-89-8104181

data of class 2 can be done by data level approach.

Therefore, hybrid data level approaches (SMOTE and bootstrap sampling) based on OVA technique with different learning methods empirically are applied to handle the imbalanced multi-class dataset. In this research, the synthesis of the new sample cluster for a small class (SMOTE) is another option or way applied to manage the classification of the imbalanced data to improve the model efficiency in predicting the data more accurately. Under sampling is the sampling technique by reducing the number of instances, which the majority class has the same amount as the minority class [32]. In this work, bootstrap technique is used as an under-sampling method. Bootstrapping approach is used to decrease the instance of the majority class group. In addition, the OVA strategy with four different classifiers (deep learning, stacking algorithm, random forest, and gradient boosting tree) is trained to evaluate and classify the models. The OVA strategy is applied to change the multi-class learning problem into a two-class learning one. The benefits of the presented method is the improvement in the efficiency of the model classification. This method can help improve the model efficiency in classifying the data more accurately in the group having a small number of samples. Additionally, it provides good efficiency for the domain with the imbalanced multi-class.

2 Background and Related Work

2.1 Elderly Health Condition

The elderly groups are people over 60 years and divided into three groups in Thailand; social bound, home bound and bed bound. Group 1- Social bound refers to the elderly who can help themselves and lead normal lives independently. Moreover, they are able to do their basic routines continuously, being in good health and having no chronic diseases or no more than two chronic diseases which can be controlled. Furthermore, they can help other people, society, community and can participate in social activities. Group 2- Home bound refers to the elderly who can help themselves, still need help from other people in some instances, have some limitations in leading their lives, have chronic diseases which cannot be cured and have both physical and mental complications resulting in doing their basic routines. Group 3- Bed bound refers to the elderly who cannot help themselves do their basic routines completely, need other people to move them, have chronic diseases which cannot be controlled and have complications, not able to help themselves or are paralyzed [27].

2.2 Imbalanced Data Problem

The main problem of the data classification is the imbalanced dataset which is caused by having two data clusters or more meaning that the data in each class is not equal. Definitely, if the data with clearly different quantity is taken through the classification method, the learning model will be classified in a majority group. There are several solutions for managing imbalanced data problems such as data level approaches, algorithm level approaches, and cost-sensitive approaches [11, 28].

2.3 Data-Level Approaches

Data-level is a technique to increase or decrease instances from the imbalanced datasets which can improve the classification accuracy. Data-level approaches play a vital role in imbalanced classification by reducing the distribution of examples and adjusting the balance of classes. These techniques are used to prepare a dataset before the classification stage. After applying datalevel solutions, the training set can learn data more efficiently [2]. Data-level solutions are separated into three categories; oversampling, undersampling and hybrid methods [11,2 8].

The oversampling technique or upsampling increase numbers of the minority class data to have a nearly equal amount of data to the majority class by using SMOTE technique [12, 19, 23, 26, 32]. SMOTE is a technique used in solving the classification of the imbalanced data because the data in each class has a different amount causing the data classification results to increase the minority class and lead to make data have more balance [14, 31]. In the random step, only one value is taken from the data values in the minority class to increase the amount of the minority class which makes the data set have more balance. In the data value random in minority class, one sample value is taken, followed by considering the K-nearest data value. Next, the Euclidean distance is calculated between the randomized data value and each nearest data value to find the least distance value. After that, the new data is synthesized to have the same value with the data giving the least distant resulting in making the amount of the data balance in every cluster.

Under sampling or down sampling is a way of reducing the majority sample group to have the same or nearly the amount of the minority sample group leading to adjusting the dataset balance before applying the training model [32]. Bootstrap sampling is a statistical technique used in resampling with replacement from the original sample. The principle of sampling technique is to increase the amount of data by sampling several clusters where each cluster will have two out of three of the previous data size. By this method, each cluster will not have the same data 100 percent of the time resulting in gaining several sampling clusters from the previous data. In this study, the bootstrap sample is used to randomize the sample only one value at a time with the total amount of n times. The values gained will return to the dataset before conducting the next sampling [5, 35].

The hybrid data level approach is the combination between oversampling and under sampling [29, 32]. It is used to increase and decease the instances of each class before the classification step. Moreover, it is the way to take the over and the under samplings to be applied to find the middle value in sampling the dataset between the two groups. This is because increasing too much data may result in causing the data bias while decreasing too much data may lose the important data in model creating. Hence, this method can help handle the imbalanced multi-class.

2.4 Multi-Class Classification.

The multi-class classification method is the way to classify the data set of more than two classes via changing the problem and dividing the classification into two types. Generally, the popular method of classifying the multi-class into two classes is the OVA approach which is the method of reducing the learning technique from the multi-class learning problem to the two-class learning problem. The OVA method is classified as a K- binary classifier. The classier is built for each class so the classifier is trained by the C_i class and the total instance groups of other classes. The result from the binary classification can make a decision by using this function: F(x) = argmaxi=1…K $f_i(x)$ determining to choose the highest value with the sample tested by OVA method; K-binary classifier. The classifier is built for each class so the classifier is trained by the C_i class and the all-sample group of other classes [7]. The advantage of OVA strategy is the fast-processing time consumption. Thus, this study chooses the OVA technique applied with different learning methods to reduce the processing complexity of the classification model and improve the model efficiency in classifying the data to be more accurate [15, 18].

2.5 Ensemble Method

Ensemble method is the technique focusing on improving the accuracy of the simulated model result using several classification models to help find the answer of which the combination of the models can increase the high accuracy of the results. This method tends to take only one training dataset [16]. However, to make the model have more functions, the model is created by using different classification techniques. After creating the ensemble models, the models are taken to predict the new data. Due to the ensemble model having several models, each model will give its result and all results will be considered to have the most appropriate answers through voting [17, 34].

2.6 The Performance Assessments of Classifiers are as Followed

Accuracy is one metric measurement for classification models.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

Where TP = the number of true positives, TN = the number of true negatives, FP = the number of false positives, and FN = the number of false negatives [15].

Sensitivity or true positive rate is the proportion of instances where it is predicted as a positive label [15].

$$SN = \frac{TP}{TP+FN} \tag{2}$$

Specificity or true negative rate is the proportion of the instances where it is predicted as a negative label [15].

$$SP = \frac{TN}{TN+FP} \tag{3}$$

G-means (geometric mean) is the measurement performance of the class-imbalanced classifiers. It is applied to balance the sensitivity and the specificity values. If the sensitivity and the specificity are equal, G-means has the maximum value [15].

$$G - mean = \sqrt{(SN * SP)} \tag{4}$$

2.7 Related Work

Other researchers have presented a combination of data level techniques using SMOTE and under sampling. Pristyanto, et al., [21] presented the data level approach to balance the class distribution on educational data mining (EDM). The combination of the SMOTE and the OSS techniques were used to handle the imbalance on educational data mining. The OSS approach is applied to split the majority class into four sections; noise sample, borderline sample, redundancy sample, and safety sample. The results indicated that the hybrid of SMOTE and One-Sided Selection (OSS) provided the best effectiveness when using SVM as a classifier.

The hybrid simulated annealing approach (SA) with different classifiers (discriminant analysis, SVM, decision tree, and KNN) was presented by Desuky et al., [5]. The SA method is applied to manage the unbalanced data and to select majority examples on UCI and KEEL datasets. Furthermore, the F-score metric with objective function is used to choose instances. The outputs of experiment illustrated that the SA method with 4 different classifiers received the best accurate rate on 9 binary datasets.

Kou et al., [13] combined the methods of resampling and ensemble model for the credit and the finance evaluation data. In this work, the resampling technique was applied to the clustering and distance-based imbalance learning mode or called the CDEILM. Moreover, the cluster-size-based resampling model was developed to divide the group size of the under sampling rate in the resampling step. The outputs of a hybrid approach illustrated that the effectiveness of AUC and G-measure values were higher than other approaches.

In addition, this method could help solve the problems of the domains in the finance imbalanced datasets.

Zhaozhao et al., [33] presented the method of the hybrid SMOTE and k-means for the imbalanced medical data. The cluster-based oversampling algorithm was applied to identify the class of all instances. The results of this paper indicated that the sensitivity and the specificity values were at 99.84% and at 99.56% when the random forest algorithm is used to classify the model. Astha, et al., [1] proposed the method of the SMOTE and the cluster-based undersampling for the imbalanced multiclass. This approach was applied to balance and preprocess the training sets. The SCUT method received higher accuracy than other methods when applied for the preprocessing data.

The combination SMOTE and undersampling method for the imbalanced datasets was proposed by Hanskunatai [9]. The hybrid sampling technique between SMOTE and undersampling technique by using the decision tree and naïve bayes was the data classifier. The DBSCAN algorithm is used to divide a group of positive and negative classes. Negative class is removed 50% with this algorithm. The results demonstrated that the hybrid sampling technique obtained the F-measure more than other sampling approaches in 7 datasets (Haberman, Glass6, glass0, vehicle2, new-thyroid1, new-thyroid2, and yeast3) when decision tree was used as the classifier. The outputs indicated this method could help improve performance of predictive model. Furthermore, the proposed method achieved the highest F-measure in Wiscosin at 0.962. Additionally, Xu et al., [23] presented a technique of the new hybrid SMOTE, Tomek-link and combined cleaning and resampling (CCR-SMOTETL). In this work, the CCR-SMOTETL approach was used to random the instances and to detect the noise data. The research results demonstrated that the effectiveness of the classification received better accuracy than other methods when classified using Random Forest algorithm.

3 Materials and Methods

3.1 Dataset

The dataset of the elderly health condition survey used in this experiment was received from Maeka Health Care Promotion Hospital, Muang District, Phayao Province, Thailand. The data was collected from January, 2022 to February 2022 in which the details of the dataset is shown in Table 1. The elderly health condition dataset is divided into three different classes; social bound (class 0), home bound (class 1), and bed bound (class 2). An example of the dataset for the experiment is shown in Figure 1. The descriptions of each feature for the elderly health condition data are presented in Table 2 in supplementary materials. Furthermore, the names of the data level approach and the associated abbreviations are shown in Table 3.

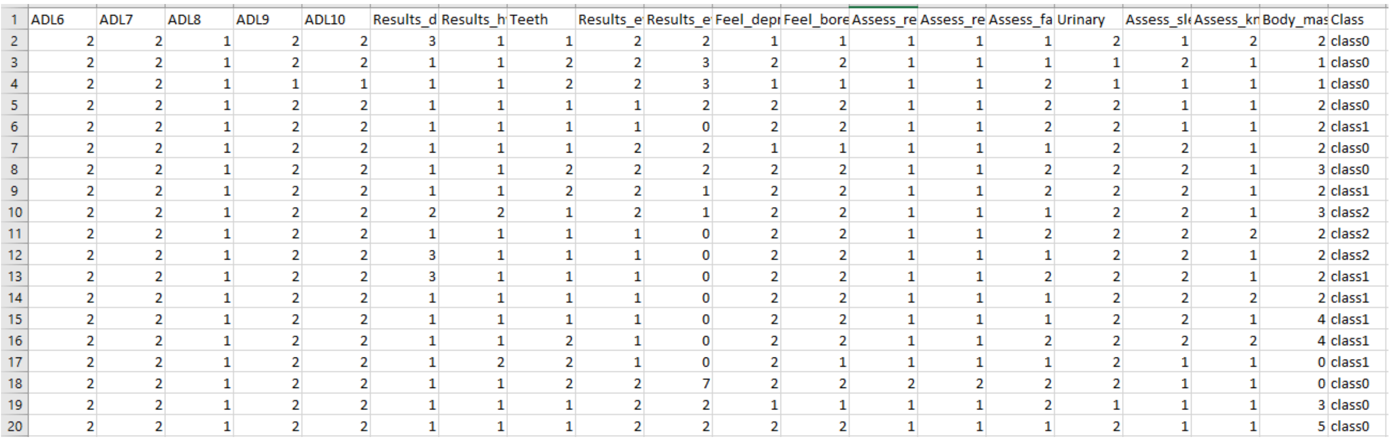
Table 1: Showing the details of the data feature used in the experiment

Dataset	No. of features	No. of instances	No. of instances/ Class	Imbalanced ratio (IR)
Elderly health condition	34	1194	610/529/55	11.09

Table 2: The explanation of the value substitution of data in classification

Feature	Descriptions	Feature	Descriptions
1. Occ (Occupation)	1= Agriculture, 2=General employee, 3=Merchant/ Personal business, 4=Animal keeper, 5= Government pensioner, 6=Unemployment, 7= Others	2. Marital Status	1= Single 2= Married 3= Widowed 4=Divorced
3. BP1 (Blood pressure check, first time)	1=normal (not more than 120/80), 2=starting high (120-139/80-89), 3=high (140-159/90-99), 4=very high (>160/100) 0= no information	4. BP2 (Blood pressure check, second time)	1=normal (not more than 120/80), 2=starting high (120-139/80-89), 3=high (140-159/90-99), 4=very high (>160/100) 0= no information
5. Cdis (Having chronic disease)	1=yes, 2=No	6. Phydis (Physical disability)	1=yes, 2=No
7. Mact (Moderate movement activity)	1=Having regularly at least 3 times a week 2=Having irregularly less than 3 times a week 3= No having due to disability	8. Teeth (The present having good teeth at least 20 teeth)	1=yes, 0= No
9. Smoking	1= Never smoking, 2= Ever smoking but not smoking now, 3= Smoking	10. Alcohol Drinking	1 = Never drinking 2= Ever drinking but stopping now 3= Still drinking
11. Eye sight problem	1=yes, 0= No	12. Eye Results (The result of eye sight examination)	1=short-sighted eye, 2=long-sighted eye, 3=cataract, 4=glaucoma, 5= pterygium, 6= macular degeneration

13.	Feeling_ unpleasant	1=yes, 0= No	14.	Feeling _depressed	1=yes, 0= No
15.	Assess_fall (The assessment of falling condition)	1= less than 30 seconds 2=more than 30 seconds 3= unable to walk	16.	Urinary (The screening of urinary incontinence)	1=yes, 0= No
17.	Assess remem1 (The assessment of the dementia condition by remembering and telling all words of things correctly.	1= true 2= false	18.	Assess_remem2 (The assessment of the dementia condition by telling one's own name and age correctly)	1=true 2= false
19.	Results_of yearly diabetes check up	1=normal 2= abnormal 3= not checking	20.	Results of yearly hypertension check up	1=normal 2= normal 3= not checking
21.	Assess_sleep (Sleep problem diagnosis)	1=Have, 0= No have	22.	Assess_knee (Knee osteoarthritis diagnosis)	1=yes, 0= No
23.	ADL1 (Feeding)	0= unable, 1= needs help cutting, spreading butter 2=independent (food provided within reach)	24.	ADL2 (Grooming)	0 = needs help with personal care 1 = independent face/hair/teeth/shaving (implements provided)
25.	ADL3 (Transfer)	0 = unable – no sitting balance 1 = major help (one or two people, physical), can sit 2 = minor help (verbal or physical) 3 = independent	26.	ADL4 (Toilet use)	0 = dependent 1 = needs some help, but can do something alone 2 = independent (on and off, dressing, wiping)
27.	ADL5 (Mobility)	0 = immobile 1=wheelchair independent, including corners 2 = walks with help of one person (verbal or physical) 3 = independent	28.	ADL6 (Dressing)	0 = dependent 1 = needs help, but can do about half unaided 2 = independent (including buttons, zips, laces)
29.	ADL7 (Stairs)	0 = unable 1 = needs help (verbal, physical, carrying aid) 2 = independent up and down	30.	ADL8 (Bathing)	0 = dependent, 1 = independent (or in shower)
31.	ADL9 (Bowels)	0 = Incontinent (or needs to be given enema) 1 = occasional accident (once/week) 2 = continent	32.	ADL10 (Bladder)	0 = incontinent, or catheterized and unable to manage 1 = occasional accident (max. once per 24 hours) 2 = continent (for over 7 days)
33.	BMI (body mass index (nutritional status))	1=obese range 2=healthy weight range 3=overweight range 4=obese range 5=very obese range 0=Unknown	34.	Class (Group of elderly which divided by their abilities in doing their daily lives)	Class0= Social Bound Class1= Home Bound Class2= Bed Bound



1	ADL6	ADL7	ADL8	ADL9	ADL10	Results_d	Results_h	Teeth	Results_e	Results_e	Feel_depr	Feel_bore	Assess_re	Assess_re	Assess_fa	Urinary	Assess_sl	Assess_kn	Body_mas	Class
2	2	2	1	2	2	3	1	1	2	2	1	1	1	1	1	2	1	2	2	class0
3	2	2	1	2	2	1	1	2	2	3	2	2	1	1	1	1	2	1	1	class0
4	2	2	1	1	1	1	1	2	2	3	1	1	1	1	2	1	1	1	1	class0
5	2	2	1	2	2	1	1	1	1	2	2	2	1	1	2	2	1	1	2	class0
6	2	2	1	2	2	1	1	1	1	0	2	2	1	1	2	2	1	1	2	class1
7	2	2	1	2	2	1	1	1	2	2	1	1	1	1	1	2	2	1	2	class0
8	2	2	1	2	2	1	1	2	2	2	2	2	1	1	2	2	2	1	3	class0
9	2	2	1	2	2	1	1	2	2	1	2	2	1	1	2	2	2	1	2	class1
10	2	2	1	2	2	2	2	1	2	1	2	2	1	1	1	2	2	1	3	class2
11	2	2	1	2	2	1	1	1	1	0	2	2	1	1	2	2	2	2	2	class2
12	2	2	1	2	2	3	1	1	1	0	2	2	1	1	1	2	2	1	2	class2
13	2	2	1	2	2	3	1	1	1	0	2	2	1	1	2	2	2	1	2	class1
14	2	2	1	2	2	1	1	1	1	0	2	2	1	1	1	2	2	2	2	class1
15	2	2	1	2	2	1	1	1	1	0	2	2	1	1	1	2	2	1	4	class1
16	2	2	1	2	2	1	1	2	1	0	2	2	1	1	2	2	2	2	4	class1
17	2	2	1	2	2	1	2	2	1	0	2	1	1	1	1	2	1	1	0	class1
18	2	2	1	2	2	1	1	2	2	7	2	2	2	2	2	2	1	1	0	class0
19	2	2	1	2	2	1	1	1	2	2	1	1	1	1	2	1	1	1	3	class0
20	2	2	1	2	2	1	1	1	2	2	2	2	1	1	1	2	1	1	5	class0

Figure 1: The example of the elderly health condition dataset

Table 3: Data level approach used in the present research and the associated abbreviations

Algorithm	Abbreviation
SMOTE based on OVA strategy	SMOTE_OVA
Hybrid SMOTE and bootstrap sampling based on OVA with different learning methods (the size of the examples is 40% of majority class)	HySM_BT40%*
Hybrid SMOTE and bootstrap sampling based on OVA with different learning methods (the size of the examples is 50% of majority class)	HySM_BT50%*
Hybrid SMOTE and bootstrap sampling based on OVA with different learning methods (the size of the examples is 60% of majority class)	HySM_BT60%*

* The proposed method

3.2 Proposed Method

In this study, the researcher presents the method of the imbalance multiclass management by applying SMOTE and bootstrap techniques based on an ensemble learning method shown in the research conceptual framework as in Figure 2. The processes of the proposed method are as follows.

Step 1: Data collection

This part involves data collection on the elderly's health condition survey. Then, the data is kept in a MySQL database and changed into a needed file, a CSV file, easy to be applied.

Step 2: Data preparation

The method of the missing data substitution has many techniques which are statistical method or mining technique used for substituting the missing data, for example, listwise data deletion [10] is an easy technique in managing the missing data. This approach is done by analyzing only the complete data. It is suitable when having a small amount of missing data and the synthesis result is very clear which is mainly applied by default. For presentation, the substitution technique of missing data is used with the unknown value.

Mean substitution is a technique of replacing the missing data by the value-known data mean in each small group of the variables [10]. It is used due to the hypothesis that the value of the missing data should be relied on the sample unit feature which should have the same interesting data value. Moreover, the class instances of unknown values will be deleted. In this case, the data has many lost features, but not more than 5 percent of all data through deleting the instances having missing data.

Data transformation is the method of converting, extracting and mapping data into a usable format. In this research, the discretization technique is used to convert numerical to nominal data. Discretization is a method of changing the number data value into the data with a small size data. This method is done before the data preparation step which reduces the data process through decreasing the size and the data complexity [8]. The steps of discretization value used in clustering the features is determined by the user and the expert. In addition, the attributes are discretized by claiming the values from the laboratory.

Step 3: Hybrid data level approaches using SMOTE and bootstrap sampling technique

Data preprocessing is to prepare the data processing and managing the data before classification. It is the data preparation which is an important process of the machine learning. If the data preparation is not done well, it will result in the operating efficiency of the other processes. In this step, the problem of the dataset having different spreading classes is managed by adjusting the class balance to have close or an equal number through the combination of SMOTE and bootstrap techniques in solving the imbalanced problems. SMOTE helps synthesize the minority class to have the same size as the majority class while the undersampling approach reduces the sample group with the majority class having the same number

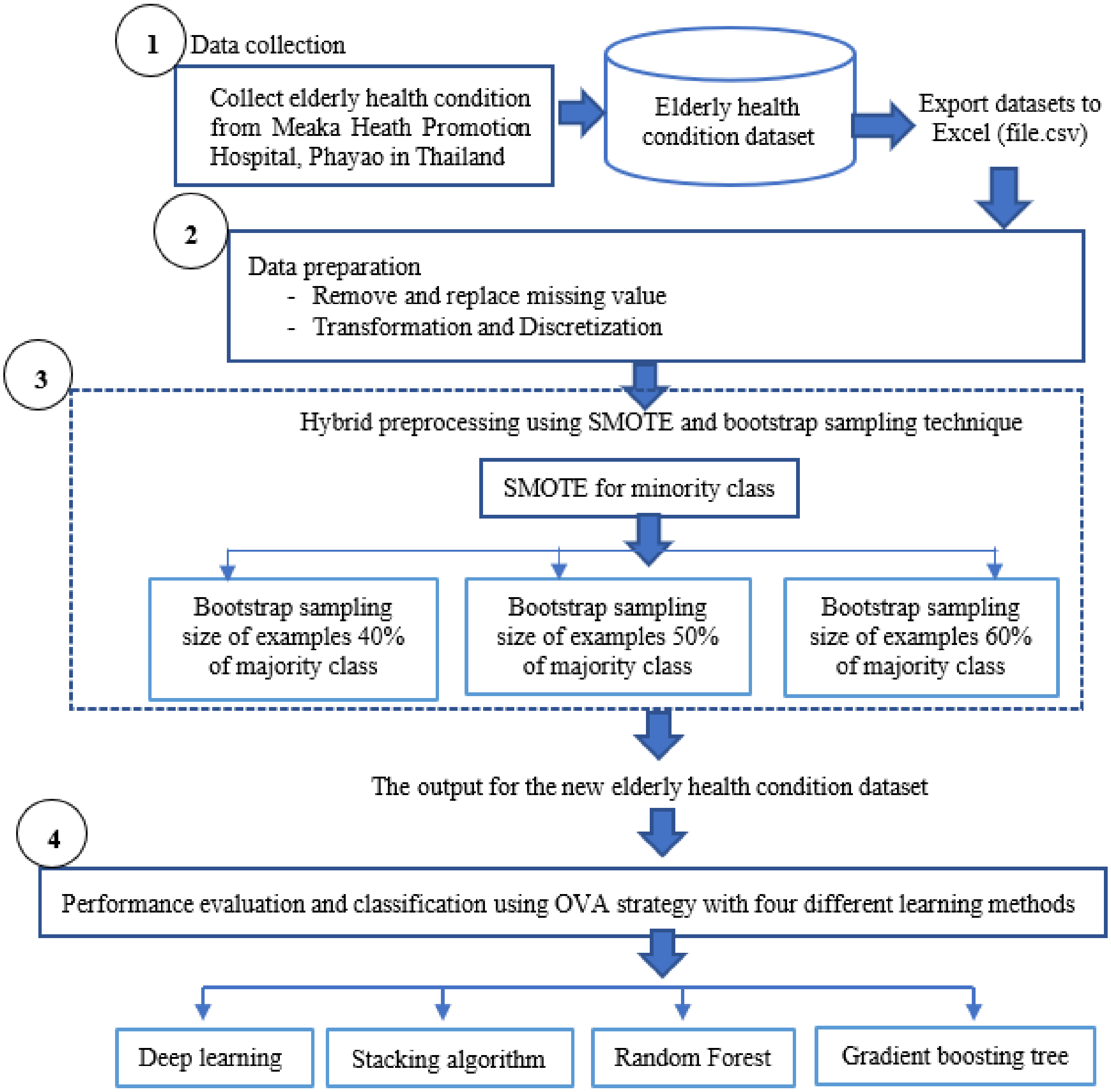

1
Data collection
Collect elderly health condition from Meaka Heath Promotion Hospital, Phayao in Thailand
Elderly health condition dataset
Export datasets to Excel (file.csv)
2
Data preparation
- Remove and replace missing value
- Transformation and Discretization
3
Hybrid preprocessing using SMOTE and bootstrap sampling technique
SMOTE for minority class
Bootstrap sampling size of examples 40% of majority class
Bootstrap sampling size of examples 50% of majority class
Bootstrap sampling size of examples 60% of majority class
The output for the new elderly health condition dataset
4
Performance evaluation and classification using OVA strategy with four different learning methods
Deep learning
Stacking algorithm
Random Forest
Gradient boosting tree

Figure 2: Framework of hybrid data level approaches based on stacking learning for imbalanced classification

or size as the minority class sample. In this research, bootstrap sampling is used in randomizing the data sample in order to make each class balanced before creating the data classification. The bootstrap sampling uses the operator to select the instances according to the sample size. The best configuration of parameter sample size is found by optimizing the parameters process using the RapidMiner software with approximately 40% - 60% of the majority class.

Step 4: Performance evaluation using deep learning and ensemble classifiers for multi-class classification

This research has improved the efficiency of the multi-class classification by using OVA strategy to divide the multi-class problem into two class or binary class together. After adjusting the dataset via the combination SMOTE and bootstrap technique, the new dataset is classified into a cluster by using OVA technique to define the main class as a positive class and the rest of other classes to be a negative class. Then, the new dataset is led to create a model to classify the data type by applying four learning classifiers. In this step, four learning methods; deep learning, stacking algorithm, random forest, and gradient boosting tree are used for the classification and evaluation models [30].

- Deep learning (DL) [20, 22, 25] is a category of neural network model as a multilayer perceptron (MLP). It is used to train with learning algorithms. Deep learning as supervised learning can help seek patterns from a large data and create models for model prediction.
- Stacking algorithm (SK) is a type of the ensemble learning method. It is one of the most popular ensemble classifiers.

This method typically creates a heterogeneous ensemble. It can combine the predictions from multiple classifies. Stacking method is used to train multiple models for solving similar problems [25, 34].

- Random forest method (RF) [4, 34] is an ensemble decision tree. By the principle, the random forest trains the same model for several times or instances on the same data set. Each time the training will choose a different trained data. Then all model decision making is voted upon. The advantage of this algorithm is to consume the rapid time in training the model. Furthermore, it avoids the risk of overfitting.
- Gradient boosting tree (GBT) is an ensemble method that helps improve the accuracy of trees. It can train either regression or classification tree models. The GBT learning algorithm is similar to a boosting technique and a decision tree. The principle of GBT is that the sampling and creating of the decision tree from different simulations and each simulation is assessed until gaining the complete decision tree model. The principle of the GBT method is to create multiple decision tree models and to evaluate each model of a decision tree. This algorithm provides a complete decision tree [4].

4 Results

In this section we will present the results of the experiment and discussion through comparing combinations of SMOTE and undersampling technique based on different classifiers (deep leaning, stacking algorithm, random forest, and gradient boosting tree) and single sampling techniques. It is compared with the accuracy rate, sensitivity, specificity, and G-means in the elderly health condition dataset.

For Table 4, it is the comparison of the effectiveness of accuracy on classification data via hybrid data level approaches; SMOTE and bootstrap technique, based on OVA strategy with different learning methods. From Table 4, it shows the effectiveness of the model classification through the HySM_BT50% method by using RF algorithm gaining the highest accuracy value at 90.11% as well as the HySM_BT60% when using RF method as classifiers at 88.48%. For the HySM_BT40% technique using SK algorithm has good accuracy of classification at 88.13%. In Table 4, it is clearly seen that the output of the HySM_BT50% method obtained is the best at an average accuracy rate of 88.01%.

Moreover, the HySM_BT50% using DL method is equal to that of the HySM_BT50% by GBT algorithm (87.18%). On the other hand, the SMOTE_OVA technique learns better than HySM_BT40% and HySM_BT60% approaches when using GBT as classifier. In summary, the performance of classification is superior to the other methods when the imbalanced data management is applied by using the hybrid data level approach based on OVA strategy with four different classifiers.

Tables 5 and 6, show the comparison of the classification efficiency in terms of sensitivity and specificity values. From Table 5, the result indicates the highest sensitivity gain is at 0.9011 when combining SMOTE and bootstrap sampling technique based on OVA with RF classifiers. Also, in Table 6, the specificity value is classified by RF algorithm which gains the highest value at 0.9480.

The result from Table 7 indicates that the HySM_BT50% by using RF classifier to assess the model efficiency which shows the G-means value gaining higher than other methods. Likewise, the HySM_BT40% with SK and HySM_BT60% with RF algorithm obtains the high G-means. The HySM_BT50% method gives the best G-means value at 0.9242.

The implementation of hybrid data level methods uses SMOTE and bootstrap technique. The chart in Figure 3 compares the distribution of samples of each class after resampling data. There are 3 methods consisting of HySM_BT40%, HySM_BT50%, and HySM_BT60%. Each class represents a group of the elderly. Class_High or class 0 refers to a group of the elderly which are social bound; Class_Middle or class 1 refers to a group of elderly that are home bound; and Class_low or class 2 refers to a group of elderly that are bed bound.

5 Discussion

When considering the efficiency of the hybrid sampling technique together with one versus group, it shows that the technique presented gives the highest correctness value when classifying the data by ensemble learning (RF). For example, the HySM_BT50% method achieves the maximum average accuracy using four different learning methods (DL, SK, RF,

Table 4: Performance of classification accuracy (%) using hybrid data level approaches based on four different learning methods

Method	No. of Instances	DL	SK	RF	GBT	Average
Original dataset	1194	71.51	72.91	75.14	72.35	72.98
SMOTE_OVA	1830	85.06	84.52	84.88	85.79	85.06
Bootstrap	165	83.33	83.33	81.25	77.08	81.25
HySM_BT40%*	732	85.39	**88.13**	87.67	84.47	86.42
HySM_BT50%*	915	**87.18**	87.55	**90.11**	**87.18**	**88.01**
HySM_BT60%*	1098	85.15	84.24	88.48	83.94	85.45

Table 5: Comparison of sensitivity using hybrid data level approaches based on four different learning methods

Method	DL	SK	RF	GBT
Original dataset	0.7151	0.7291	0.7514	0.7235
SMOTE_OVA	0.8506	0.8452	0.8488	0.8579
Bootstrap	0.8333	0.8333	0.8125	0.7708
HySM_BT40%*	0.8539	0.8813	0.8767	0.8447
HySM_BT50%*	0.8718	0.8755	0.9011	0.8718
HySM_BT60%*	0.8515	0.8424	0.8848	0.8394

Table 6: Comparison of specificity using hybrid data level approaches based on four different learning methods

Method	DL	SK	RF	GBT
Original dataset	0.8339	0.8433	0.8581	0.8395
SMOTE_OVA	0.9193	0.9161	0.9182	0.9235
Bootstrap	0.9091	0.9091	0.8966	0.8706
HySM_BT40%*	0.9212	0.9369	0.9343	0.9158
HySM_BT50%*	0.9315	0.9336	0.9480	0.9315
HySM_BT60%*	0.9198	0.9145	0.9389	0.9127

Table 7: Comparison of G-means using hybrid data level approaches based on four different learning methods

Method	DL	SK	RF	GBT
Original dataset	0.7722	0.7841	0.8030	0.7796
SMOTE_OVA	0.8843	0.8799	0.8828	0.8901
Bootstrap	0.8704	0.8704	0.8538	0.8192
HySM_BT40%*	0.8869	0.9087	0.9051	0.8796
HySM_BT50%*	0.9012	0.9041	0.9242	0.9012
HySM_BT60%*	0.8850	0.8777	0.9115	0.8753

GBT). The instances obtained from combination sampling have the middle value in the data of a majority class and a minority class.

The output of this research shows that the hybrid data-level approaches with ensemble learning method (RF and SK) can help improve the efficiency of the model. Also, the accuracy rate is increased in a balanced sample of an elderly health condition dataset. The hybrid data level methods with ensemble learning provides the diversity of models which might reduce the bias of learners and decrease the skewed distribution for the imbalanced dataset classification.

In summary, all results clearly indicate that the hybrid data-level approaches based on OVA with different learning are applied to increase the efficiency of the data classification and to manage the imbalance of the elderly health condition data. Moreover, this technique still shows that it is an appropriate approach for the dataset with low imbalance ratio for multi-class.

The research also considers the procedural similarity hybrid data-level sampling techniques. From the previous studies [21] presents the combination SMOTE and OSS technique on EDM. SVM algorithm is used to predict the model and to improve accuracy of classification. The hybrid data-level solutions can help reduce the skew of each class and provide good effectiveness of classification in EDM. The differences from the previous studies are as follows; firstly, applying the hybrid data level approaches between SMOTE and bootstrap technique for elderly health condition dataset. In this process, there are resampling of instances approximately 40%- 60% of majority class to balance each class and to reduce the skew in the class

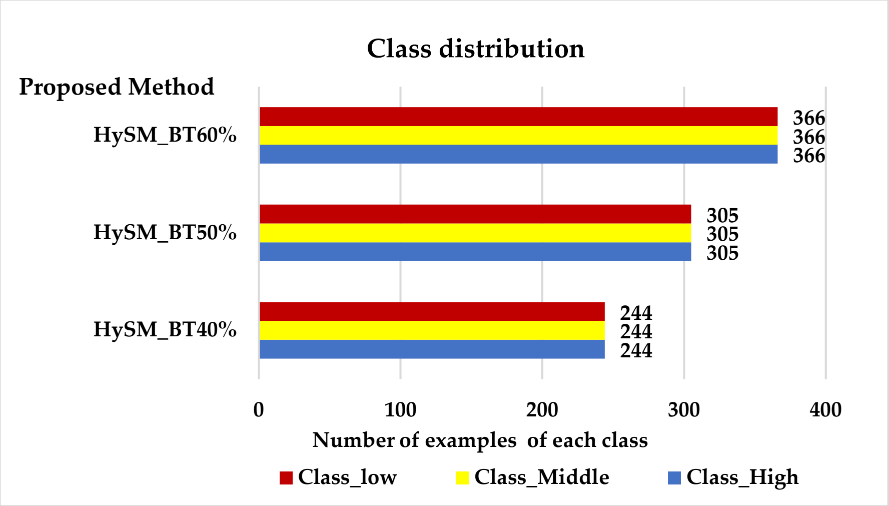


Figure 3: The distribution of instances per class using hybrid data level approach

distribution. Then, the data was applied to create the model for data classification. Secondly, in the process before evaluation of the classification model, the OVA strategy is applied to split the multi-class into binary classes. The validation splits up the example dataset into a training set (70%) and a test set (30%). Finally, four learning algorithms; deep learning, stacking algorithm, random forest, and gradient boosting tree are used in the evaluation process of the model classifier.

6 Conclusions

This study presents the hybrid method which the SMOTE and bootstrap sampling based on OVA technique with four different classifiers for the imbalanced multi-class management. The hybrid data-level approaches include HySM_BT40%, HySM_BT50%, and HySM_BT60%. The elderly's health condition data used in this research was obtained from Meaka Health Promotion Hospital in Thailand. It revealed that the efficiency of model classification has more accuracy. The HySM_BT50% method achieved the highest correctness when RF algorithm is used to classify the data. The best effective rate of classification with an accuracy of 90.11% (Sensitivity = 0.9011, Specificity = 0.9480, and G-means = 0.9242). The research results indicate that the combining sampling technique based on OVA strategy with DL, SK, RF, and GBT classifiers provides better classification accuracy rates than the single sampling approach because combination SMOTE and bootstrap sampling in the class imbalance is reduced. The advantage of the technique presented is to make the gained sample have the middle value between the data in the majority cluster and the data in the minority cluster, and helps increase the efficiency of the prediction model. This approach receives a prototype model for imbalanced multiclass handling. Additionally, the balance model is taken to create a prediction model to analyze the elderly health condition and a plan to promote elderly health care. Furthermore, the proposed approach might be a good choice for the dataset that has a large number of instances with low imbalance ratio. In the future, the feature selection method is applied in classifying the imbalanced data in order to select the feature having the importance and relating to each other. Moreover, the presented method might be applied to the other real-world problems.

Acknowledgments

This research project is supported by the Thailand science research and innovation fund and the University of Phayao (Grant No. FF66-RIM011). The Meaka Heath Promotion Hospital of Phayao in Thailand provided useful advice in data collection.

References

[1] A Agrawal, H. L. Viktor, and E. Paquet, "SCUT: Multi-Class Imbalanced Data Classification Using SMOTE and Cluster-Based Undersampling," *Proceedings of 7th International Joint Conference*, IC3K, Lisbon, Portugal, pp. 226-234, 12-14 November 2015.

[2] H. Ali, M. M. Salleh, K. Hussain, A. Ahmad, A. Ullah, A. Muhammad, and M. A. Khan, "Review on Data Preprocessing Methods for Class Imbalance Problem," *Int. J. Eng. Technol.*, 8:390-397, 2019.

[3] H. Ali, M. M. Salleh, R. Saedudin, K. Hussain, and M. F. Mshtag, " Imbalance class prolems in data mining: A Review," IJEECS, 14:1560-1571, 2019..

[4] A. Csörgő Bentéjac, and G. Martínez-Muñoz, "A Comparative Analysis of Gradient Boosting Algorithms," *Artificial Intelligence Review*, 54:1937-1967, 2021.

[5] S. Desuky and S. Hussain, "An Improved Hybrid Approach for Handling Class Imbalance Problem," *Arab J Sci Eng.*, 46:3853-3864, 2021.

[6] S. M. A. Elrahman and A. Abraham, "A Review of Class Imbalance Problem," *JNIC*, 1:332-340, 2013.

[7] E. Barrenechea Fernández, H. Bustince, and F. Herrera,

"An Overview of Ensemble Methods for Binary Classifiers in Multi-Class Problems: Experimental Study on One-vs-One and One-vs-All Schemes," *Pattern Recognition*, .44:1761-1776, 2011.

[8] S. García, J. Luengo, and F. Herrera, "Tutorial on Practical Tips of the Most Influential Data Preprocessing Algorithms in Data Mining," *Knowledge-Based Systems*, 98:1-29, 2016.

[9] A. Hanskunatai, "A New Hybrid Sampling Approach for Classification of Imbalanced Datasets," *Proceedings of 2018 3rd International Conference on Computer and Communication Systems (ICCCS)*, IEEE. Nagoya, Japan, pp. 67-71, 27-30 April 2018.

[10] R. Houari, A. Bounceur, A. K. Tari, and M. T. Kecha, "Handling Missing Data Problems with Sampling Methods," *Proceedings of International Conference on Advanced Networking Distributed Systems and Applications*, IEEE, Bejaia, Algeria, pp. 99-104, 17-19 June 2014.

[11] P. Kaur and A. Gosain, "Robust Hybrid Data-Level Sampling Approach to Handle Imbalanced Data During Classification," *Soft Computing,* 24:15715-15732, 2020.

[12] S. Kotsiantis, D. Kanellopoulos and P. Pintelas, "Handling Imbalanced Datasets: A Review," *GESTS International Transaction on Computer Science and Engineering*, 30:25-36, 2006.

[13] G. Kou, H. Chen, and M. A. Hefni, "Improved Hybrid Resampling and Ensemble Model for Imbalance Learning and Credit Evaluation," *JMSE*, 7:511-529, 2022.

[14] W. J. Lin, and J. J. Chen, "Class-Imbalanced Classifiers for High-Dimensional Data," *Briefings in Bioinformatics*, 14:13-26, 2013.

[15] C. Lorena, A. C. De Carvalho, and J. M. Gama, "A Review on the Combination of Binary Classifiers in Multiclass Problems," *Artificial Intelligence Review*, 30:19-37, 2009.

[16] B. Mahesh, "Machine Learning Algorithms-A Review, *Int. J. Sci. Res*., 9:81-386, 2020.

[17] R. Kora Mohammed, "A Comprehensive Review on Ensemble Deep Learning: Opportunities and Challenges," *Journal of King Saud University-Computer and Information Sciences*, 35:757-774, 2023.

[18] T. H. Oong and N. A. M. Isa, "One-Against-All Ensemble for Multiclass Pattern Classification," *Appl. Soft Comput.*, 12:1303-1308, 2012.

[19] K. Polat Ozdemir and A. Alhudhaif, "Classification of Imbalanced Hyperspectral Images using SMOTE-Based Deep Learning Methods," *Expert Syst. Appl.*, 178:114986, 2011.

[20] Y. Pandey, "Credit Card Fraud Detection using Deep Learning," *IJARCSSE*, 8:18-25, 2017.

[21] Y. Pristyanto, I. Pratama, and A. F. Nugraha, "Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification," *Proceedings of 2018 International Conference on Information and Communications Technology,* IEEE*,* Yogyakarta, Indonesia, pp. 310-314, 06-07 March 2018.

[22] A. Pumsirirat and Y. Liu, "Credit Card Fraud Detection Using Deep Learning Based on Auto-Encoder and Restricted Boltzmann Machine," *Int J Adv Comput Sci Appl*, 9:18-25, 2018.

[23] E. Rendón, R. Alejo, C. Castorena, F. J. Isidro-Ortega, and E. E. Granda-Gutiérrez, "Data Sampling Methods to Deal with the Big Data Multi-Class Imbalance Problem," *Appl. Sci.*, 10:1-15, 2020.

[24] J. A. Sáez, B. Krawczyk, and M. Wozniak, "Analyzing the Oversampling of Different Classes and Types of Examples in Multi-Class Imbalanced Datasets," *Pattern Recognition*, 57:164–178, 2016.

[25] A. Shrestha and A. Mahmood, "Review of Deep Learning Algorithms and Architectures," *IEEE Access*, 7:53040-53065, 2019.

[26] Y. Sun, A. K. C. Wong and M. S. Kamel, "Classification of Imbalanced Data: A Review," *Int. J. Pattern Recognit. Artif.*, 23:687-719, 2009.

[27] C. Supromin and S. Choonhakhlai, "The Provision of Public Services in Municipalities in Thailand to Improve the Quality of Life of Elderly People," *Kasetsart Journal of Social Sciences*, 40:619-627, 2019.

[28] K. Upadhyay, Prabhjot Kakur, and S. Prasad, "A Review on Data Level Approaches to Address the Class Imbalance Problem," *Proceedings of International Conference on Recent Challenges in Engineering Science and Technology*, Andhra Pradesh, India, pp. 152-158, 9-10 April 2021.

[29] K. Upadhyay, P. Kaur, and D. K. Verma, "Evaluating the Performance of Data Level Methods using Keel Tool to Address Class Imbalance Problem," *Arab J Sci Eng*., pp. 1-14, 2022.

[30] S. Wan and H. Yang, "Comparison Among Methods of Ensemble Learning," *Proceedings of 2013 International Symposium on Biometrics and Security Technologies,* IEEE, Sichuan, China, pp. 286-290, 2-5 July 2013.

[31] Z. Xiang, Y. Su, J. Lan, D. Li, Y. Hu, and Z. Li, "An Improved SMOTE Algorithm Using Clustering," *Proceedings of 2020 Chinese Automation Congress (CAC) IEEE*, Shanghai, China, pp. 1986-1991, 06-08 November 2020.

[32] B. Xu, W. Wang, R.Yang, and Q. Han, "An Improved Unbalanced Data Classification Method Based on Hybrid Sampling Approach," *Proceedings of 2021 IEEE 4th International Conference on Big Data and Artificial Intelligence (BDAI),* IEEE, Qingdao, China, pp. 125-129, 2-4 July 2021.

[33] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A Cluster-Based Oversampling Algorithm Combining SMOTE and k-Means for Imbalanced Medical Data," *Information Sciences*, 572:574-589, 2021.

[34] Y. Zhang, J. Liu, and W. Shen, "A Review of Ensemble Learning Algorithms Used in Remote Sensing Applications," *Applied Sciences*, 12*:*1-20, 2022.

[35] Y. Zhao, Z. S. Y. Wong, and K. L. Tsui, "A Framework of Rebalancing Imbalanced Healthcare Data for Rare Events' Classification: A Case of Look-Alike Sound-Alike Mix-Up Incident Detection," *J. Healthc Eng.*, pp. 1-11, 2018.

Rattanawadee Panthong (photo not available) received the Ph.D. (Computer Science) from Kasetsart University, Thailand in 2021. She is currently an instruction at the School of Information and Communication Technology at the University of Phayao, Thailand. In addition, she is serving as an Assistant Dean in the School of Information and Communication Technology (2019–present). Her research interests are in machine leaning, data mining, data analytics and data warehouse. She can be contacted at email: rattanawadee.pa@up.ac.th.

End-to-End Open-Domain Question-Answering System: Baseline and Case Study using EIAD Dataset

Moataz Mohammed*, Salsabil A. El-Regaily, and Mostafa M. Aref
Ain Shams University, Abbassiya, Cairo, EGYPT

Abstract

During the artificial intelligence (AI) era, AI has evolved into a multidisciplinary industry in all domains. (NLP) Natural language processing, represents one of the most fascinating AI tasks. It can perform a number of tasks including question answering (QA), machine translation, entity linking, text generation, topic modelling, text summarization, and text to speech (TTS). The QA task is the focus of this research. It highlights the Open-Domain Question-Answering ODQA task explained using the field of Islamic religion. In this research, the QA task presents a model for developing an IslamBot QA system. IslamBot is a question-and-answer chatbot that is free-formed which can answer Islamic-related questions. The models of deep learning-based retrieval-reader were used to create the ODQA model. This paper uses a model that is based on data derived from the English Islamic Articles Database (EIAD). The EIAD dataset is a labelled ODQA dataset that was crowdsourced. The EIAD dataset contains approximately 10k articles, 7.5k of which were crowdsourced, and approximately 10k question-answer pairs. Every article contains at least one question-and-answer pair. This paper develops an end-to-end ODQA model that uses the EIAD dataset to create a benchmark and an entirely novel baseline model. It also sets a new standard with the most recent Dense Passage Retriever models, which achieve 78% R@100. The ODQA model also generated novel results. It received a 71.5% EM and a 75.8% F1 score. Furthermore, due to the length of the answer, the use of the long-form open-domain type is a hard issue: justification answer. Besides, the input of the model is only the question without context.

Key Words: Open-domain question answering; natural language processing; information retrieval; reading comprehension; retriever-reader.

1 1 Introduction

It has become indispensable to mimic human behavior during the past few decades. Computer Vision (CV) and Natural Answering (QA), Text Generation, Part of Speech (POS), Language Processing (NLP) are two subfields of Artificial Intelligence (AI) that mimic humans' vision and language. NLP is the subject of this study. A wide range of human language tasks can be performed using NLP. It is capable of performing many human language tasks such as: Text Summarization, Question Machine Translation, Named Entity Recognition (NER), and Text-to-Speech (TTS). The QA task is the focus of this research. It entails one person asking a question and another responding to it. It is possible to achieve this in the machine world by instructing the computer to mimic the responsible person for giving the answer [1]. The QA task can be done using NLP applications such as chatbots, which allow you to ask a question and receive an instant response from the developed chatbot. Chatbots can be created in a variety of fields, including economics, advertising, tourism, politics, ticket booking, social media, learning, call centers, industry, and religion. This study's use case is QA in the religious field. When looking for the possibility of using chatbots in Islamic websites, it was noticed that these websites are either knowledge-based or human-based. Human-based chatbots like Islam-Religion and Islam-Portal, are accessible twenty-four hours a day, seven days a week to answer any question. And as for the knowledge-based chatbots, it relies upon concepts such as knowledge graphs and decision trees, which can be found on websites such as Allah's Word, Islam House, Ask-A-Muslim, and Guide To Islam. Knowledge-based chatbots rely on a list of generic questions to select from until it finds the closest question to answer, but free-form chatbots are not available for these religious websites. Traditional QA systems are either closed-domain QA (CDQA) or reading comprehension (RC). In order to extract the answer from the RC systems, the user must provide the question and some context. However, progress proceeds in profound learning and the use of attention and processors. Systems for open-domain quality assurance (ODQA) has emerged. We can train a deep learning model on a large number of documents using open-domain QA systems. The model can then be completed by simply typing the question as input into the model without any context. This study makes a contribution by fine-tuning a long-form or free-form open-domain QA (ODQA) model on an Islamic religion dataset. The recent ODQA systems are either retriever-generator-

*Computer Science Department, Faculty of Computer and Information Sciences. Email: moataz.mohammed@cis.asu.edu.eg.

based or retriever-reader.

Furthermore, the data sets used by these systems can range from crowd-sourced datasets like Squad2 to datasets obtained from websites structured in a question-answer format, like the ELI5 dataset extracted from the Reddit website. LFODQA is a new QA sub-task, yet Hurdles [8] ranks among the most recent works of this kind of NLP task. With the ELI5 data set, it employs the retriever-generator model. Cluster Former Model [15] achieves cutting-edge performance in open domain question answering (ODQA) using a perfect match (EM) score of 68% and the Search QA dataset. With an EM score of 38.6%, the model of BERTserini [16] is an end-to-end open-domain QA model. Open-domain QA for COVID-19 [10] is a retriever-reader-based model with an EM score of 39.16% that uses Squad2 and COVID-QA datasets. We accomplished the following during this study:

1. A new benchmark in the open-domain question answering task using EIAD dataset.
2. A new end-to-end open-domain question answering model.
3. Cutting-edge results obtained while fine-tuning some of the most recent ODQA models and DPR models on the EIAD dataset.

Section 2 discusses related works on open-domain question answering as well as some research on QA datasets related to the Islamic domain. The end-to-end ODQA system architecture and the dataset used are then thoroughly discussed in Section 3. In Section 4 we demonstrate the results of our ODQA model experiments. The results are compared to the most recent state-of-the-art ODQA models. During the discussion Section 5, there are highlights for the results and models used followed by a case study in Section 6. Finally, during the conclusion in Section 7, a quick summary of this work is obtained.

2 Related Work

Before getting into our contributions to this study, we present major innovations in question answering tasks. This section discusses cutting-edge Open-Domain Question-Answering research. We concentrate on transformer-based research, such as[9, 13, 15], because they have made significant advances in Deep Learning in the last decade.

2.1 Learning Dense Representations of Phrases at Scale

The problem of answering open-domain questions can be reframed as a phrase retrieval issue. For the first time, we assert that we can learn dense representations of phrases on our own and reach much improved results in open domain QA. We present an efficient method [9] for learning phrase representations from reading comprehension tasks under supervision. We also recommend a query-side fine-tuning strategy to aid transfer learning and reduce the gap between inference and training.

2.2 End-to-End Training of Neural Retrievers for Open-Domain Question Answering

Unsupervised pre-training with the Inverse Cloze Task and masked salient spans are followed by supervised fine-tuning using question-context pairs. This approach [13] leads to absolute gains of 2+ points over the previous best result in the top-20 retrieval accuracy on Natural Questions and TriviaQA datasets. We next explore two approaches for end-to-end training of the reader and retriever components in OpenQA models.

2.3 Cluster-Former: Clustering-based Sparse Transformer for Question Answering

Cluster-Former is a new sparse Transformer based on clustering that performs attention throughout chunked sequences. The proposed framework [15] is based on two distinct Transformer layers: the Cluster-Former Layer and the Sliding-Window one. This new design enables information integration beyond local windows, which is particularly useful for question answering (QA) tasks that depend on long-range dependencies.

3 Proposed System Architecture

The architecture of the proposed Open Domain Question Answering (ODQA) system is depicted in Figure 1. This system design is a retriever-reader paradigm that focuses on obtaining related articles and extracting the query response from these top-ranked articles. The EIAD dataset [11] was used to create this study. It is a collection of English Islamic articles. Crowdworkers used the Haystack annotation tool [4] to annotate this dataset. Each module of this design will be discussed in the parts that follow.

3.1 Database

Before moving on to the system's main components, we must first discuss the dataset that was used. The Content Table and the Indexing Table are the two main tables in the database. The Indexing Table tracks and stores the content embeddings. The Content Table displays the data from the used dataset. A Collection of English Islamic Articles During this work EIAD is the target dataset. This data set was gathered from three of the most reputable and secure Islamic websites on Internet, including IslamQA [5], Islam Religion [6], and New Muslims [12]. SQUAD is the format of the EIAD dataset. The dataset [11] appears to contain 10,000 articles divided into 15 Islamic categories. Each article has its own metadata, which includes the article title, description, rating, number of views, and date.

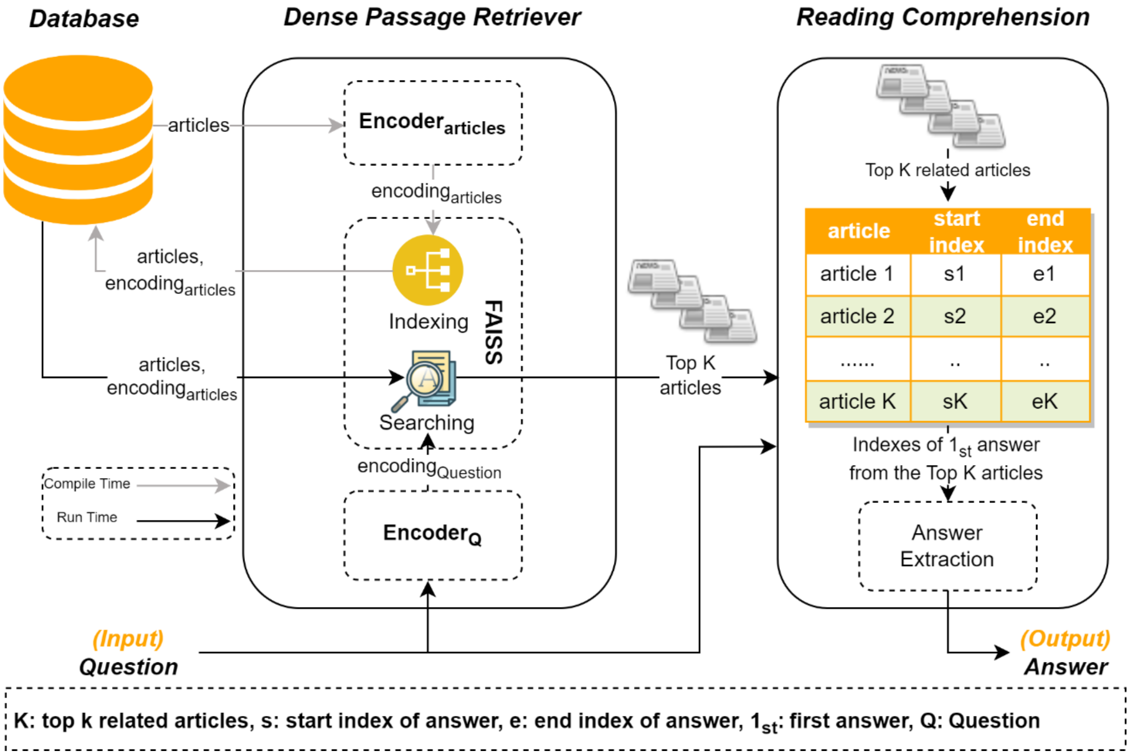
These articles were indexed using the FAISS index model [3] and stored in a SQL database. The overall number of articles in the dataset is shown in Table 1. Furthermore, the Haystack annotation tool has annotated approximately 7.5K articles. These annotated articles were used in question-answer pairs of generation 10k. The EIAD dataset contains answers to all of the questions. There is at least one answer to each question. Figure 2 shows that the length of these responses ranges from 50 to 1400 characters.

Figure 3 depicts the dataset's distribution. The EIAD dataset is divided into three subsets based on the most common distribution: training 80% with 6k annotated articles, then development 10% with 750 annotated articles, and testing 10% with another 750 annotated articles.

3.2 Retriever

Because the ODQA system is only concerned with the question, we must find the most relevant articles for this question in order to get the answer from such retrieved articles. A retrieving process is the process of finding the most relevant articles. The retriever element in Figure 1 is the ODQA system's main module. Dense Passage Retriever DPR-based [7] is the retriever here. It is equipped with a pair of encoders transformer-based [14] that serves as encoder models $E_{articles}$ and Equation. It also includes the FAISS model [3] for searching and the indexing tasks.

3.2.1 Encoder $E_{articles}$. In the dense passage retriever DPR models, Article Encoder $E_{articles}$ is the first model and component. This model functions as an encoder, taking in a text and producing a low-dimensional numerical representation vector for that text. Because this retriever model is a DPR model, the dataset that must be trainable on this DPR model must be in DPR format, which differs from the default SQUAD format. As a result, the EIAD dataset has a DPR-format replicated version. As a result, $E_{articles}$ is encoding all of the articles in the database. These encodings are then saved to be used later. They are used to select the query's most comparable K articles in the database with similar embeddings. We have more than one training trial for this model, which will be discussed further below.



article	start index	end index
article 1	s1	e1
article 2	s2	e2
......	..	..
article K	sK	eK

Figure 1: Open-domain question answering system architecture

Table 1: Dataset size

Articles	Annotated Articles	Question-Answer pairs
10k	7.5k	10k

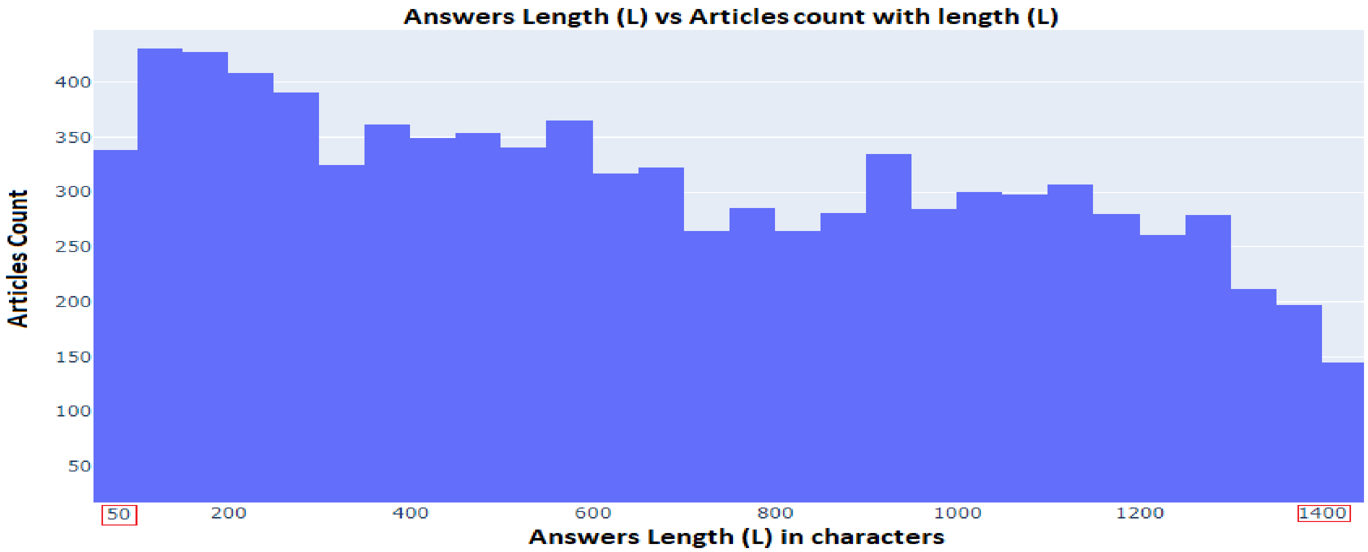


Figure 2: Articles count vs the range of answers length in characters

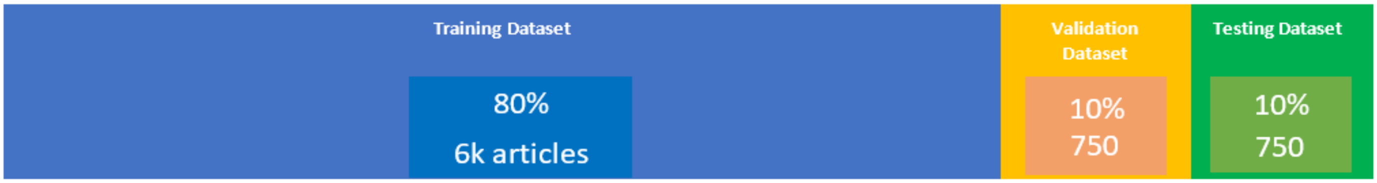


Figure 3: Dataset subset percentage

3.2.2 Encoder E_q. The articles encoder has the same concept. The question encoder model is an Eq model that encodes the asked question by taking the asked question as input text and applying the encoding process to it. As a result, it generates a low-dimensional vector for this question. This vector has a variable length and a maximum length of 50. The embedding of this question, along with the encodings of all articles, will be passed into the FAISS model later to extract the most similar articles. It is important to note that the previous models ($E_{articles}$ and E_q) are inter-correlated, which means that both had to be trained at the same time.

3.2.3 FAISS Model. Facebook created the FAISS model [3], which is a library. It is used to carry out an efficient search. The indexing procedure is what allows it to perform well in similarity searching. The indexing in the FAISS model is based on dense vector clustering. The FAISS model includes GPU and CPU support. In the following sections, we will go over the searching and indexing processes in more detail.

The FAISS flat index factory type was used to index the EIAD article dataset. The EIAD article dataset embeddings are stored with their text, like ($article_1$, $encoding_{article1}$), as a data structure of pairs, where article1 refers to the context of the first article and $encoding_{article1}$ refers to the content of the first article's encoding. So, the retriever takes the user's question and the encoder Eq encodes it in a dense vector at runtime. In addition, the retriever pulls all encodings from all the articles. The following formula is used to find the top maximum results by taking the dot product of the question vector $encoding_q$ and the vectors' articles $encoding_{articles}$. The indexes of the vector articles $encoding_{articles}$ that produced the most extreme results can then be used to obtain the top relevant articles. These indices are used to extract some other parts of pairs that include the article context.

3.3 Reader (Reading Comprehension)

The reader model's task is to extract the question's answer from the top-retrieved articles k. This was accomplished by learning how to extract the beginning and ending indexes of the answers from the original ones. Our reader is based on the Framework for Adapting Representation Models (FARM) reader. It is simple, quick, and easy to use. These readers are transformer-based, particularly the BERT [2] families. The FARM reader includes a prediction head and a built-in language model. In general, the reader is either an abstractive or an extractive reader. We use the extractive reader in our work because that domain is much more sensitive, requiring that the answer be extracted as it is.

$$articles_K = Top_K\left(DESC\left(\overrightarrow{embed_q}_{\,1*M} \cdot \begin{bmatrix} \overrightarrow{embed_{art_1}} \\ \overrightarrow{embed_{art_2}} \\ \overrightarrow{embed_{art_3}} \\ \dots \\ \dots \\ \dots \\ \overrightarrow{embed_{art_N}} \end{bmatrix}^{T}_{N*M} \right)\right) \tag{1}$$

N: Dataset size

M: Embedding length

K: The number of articles to retrieve

4 Experimental Results

The dense passage retriever is the focus of this work when building the ODQA system. The answer was extracted from the retrieved article using the reader concept. The different trials of this system's components will be discussed during this section.

During these trials, more than one model is fine-tuned. Detailed information about each model and its results can be found in the following tables. Tables 2 and 3 show the results of the reader and retriever modules, respectively. Our own EIAD dataset was used for these experiments.

The DPR model was used to begin the training, and base-bert-uncased was used to encode the question and the article. However, the result was disappointing, reporting 33% recall@100. The recall@100 improved to hit 67% in the second trial while using Facebook's context encoder and question encoder models. All-MiniLM-L6-V2 is a sentence transformer-based boosted the recall@100 to 78% which exceeded all of the prior trials. The final attempt had the greatest outcome. The configurations of these several trials are shown in Table 2, including the number of parameters in each model as well as hyperparameters such as learning rate (LR), batch size, number of epochs, and embedding dimension. Table 2 also displays the earlier outcome of these improved models. Table 3 illustrates the various experiments for the reader model. With the exception of the metrics results and model dimensions, it displays the same entries from the retriever model's table. The reader models are known as QA models that respond to questions. So, we employed the F1 and exact match EM scores as measures to evaluate the QA models. Two models—distilled-bert-base-uncased-distilled-squad and Roberta-base-squad2—were used for the majority of the trials. Based on the change in batch size and the number of epochs, the first model obtained two distinct scores for the F1 score and EM score. Using 10 batch sizes and 8 epochs, it reached 65.57% F1 score and 59.33% EM. With 16 epochs, it attained 75.8% F1 score and 71.5% EM. The second variation, Roberta-base-squad2, scored (67.7%, 67.4%) for F 1 scores and (60.5%, 61.4%) for EM. The distilled-bert-base-uncased-distilled-squad achieved the greatest results after all of these tests, with a 75.8% F1 score and 71.5% EM.

Table 2: DPR model trials

Model		**Results**	**Batch Size**	**Epochs**	**Learning Rate**	**Dimensio**	**Number Parameters**	
question	**context**	**R@100**					**question**	**context**
bert-base-uncased	bert-base-uncased	33%	16	16	3e-05	768	110M	110M
facebook/dpr-question-encoder-single-nq-base	facebook/dpr-ctx-encoder-single-nq-base	67%	8	16	3e-05	768	110M	220M
sentence-transformers/all-MiniLM-L6-v2	sentence-transformers/all-MiniLM-L6-v2	**78%**	10	16	3e-05	384	23M	23M

Table 3: Reader models trials

Model	**Results**		**Batch Size**	**Epochs**	**Learning Rate**	**Number Parameters**
Name	**Exact match (EM)**	**F1**				
distilbert-base-uncased-distilled-squad	59.33%	65.57%	10	8	1e-05	66M
roberta-base-squad2	60.5%	67.7%	10	8	1e-05	125M
roberta-base-squad2	61.4%	67.4%	32	16	1e-05	125M
distilbert-base-uncased-distilled-squad	**71.5%**	**75.8%**	10	16	1e-05	66M

Once we reached the best outcomes for the ODQA components on EIAD dataset, we can perform comparison with some of recent works in ODQA task. Comparison between our best DPR results and one of the best models in this task and the inventor of the DPR concept [7] is shown in Table 4. This comparison is constructed by fine-tuning the original DPR model on our EIAD dataset. Table 4 shows that our best DPR component trial beats the original DPR model by increasing the retrieval accuracy on the Top-100 and Top-1 by 11%. Furthermore, we contrasted one of the most current ODQA works with our Reader model (ODQA model) [16] which was fine-tuned on EIAD dataset. The EM and F1 score of this model are distinguished with those of our model in Table 5. Table 5 shows that our model performs better than [16] by 43% EM and F1 score.

5 Discussion

We have discussed an end-to-end open domain question answering ODQA task during this research. In addition, by providing the asked question without context, we demonstrate how ODQA provides a more exciting task than traditional QA. We focused on how to make an effective end-to-end ODQA system in this study, as efficiency is achieved by optimizing storage resources and GPUs used. The storage resources were optimized through one of the following ways: 1. To avoid zero values, use the Dense Passage Retriever DPR models rather than the sparse retrieval models. 2. Despite the fact that the DPR makes use of used storage, it still employs two models. As a result, we sought models that were as light as possible while maintaining accuracy. During the DPR, these models were used to compensate the model's weights of the complex ones while preserving storage optimization. The same was true for the reader model, which attempted to select a small model with excellent accuracy. GPU optimization was performed using the maximum number of batch sizes allowed for the model during fine-tuning. The maximum number of batch sizes varied from model type to model type based on model size. In spite of all this, there are some limitations with this architecture. The dimension of the retriever input

Table 4: Comparison table based on Top-1 & Top-100 retrieval accuracy of models fine-tuned on EIAD dataset

Model	**Top-1**	**Top-100**
Dense Passage Retrieval for Open-Domain Question Answering [14]	13.5	67.5
all-MiniLM-L6-V2-distilbert-base-uncased-distilled-squad (ours)	**24.4**	**78.8**

Table 5: Comparison table based on EM & F1 scores for reader models fine-tuned on EIAD dataset

Model	**EM**	**F1**
End-to-End Open-Domain Question Answering with BERTserini [4]	28.23	41.36
all-MiniLM-L6-V2-distilbert-base-uncased-distilled-squad (ours)	**71.5**	**75.8**

(question) vectors and context vectors were limited to 384 which achieved the best result. Also, the dataset must be annotated to train the model. Additionally, due to the small size of the dataset, there was an accuracy limit. The EIAD dataset contained approximately 10,000 annotated question-answer pairs. If the number of annotated question-answer pairs increased, the model's accuracy might get improved as more features were learned.

6 Case Study

This case study focuses on obtaining an answer to a religiously stated question on the Top-500 retrieved articles. As a result, it presupposes that the dataset has already been stored and collected in the database, implying that the database module portion of this case study is omitted. Furthermore, the case study focuses on exploring the process of input – output IO throughout runtime, as seen in Figure 4, which is represented by blue arrows. As a result, it illustrates the question answering system that is in charge of receiving input (query) and producing output (answer). In the example study, the Retriever-Reader architecture relies solely on the query as an input. The case study's question q is **"Why there are heaven and hell?"** because the EIAD dataset is an Islamic religion one. Encoder$_q$, a fine-tuned Dense Passage Retriever DPR model, is used to encode this query. Then, as seen in Figure 5, Encoder$_q$ outputs the embedding vector of q as E$_{question}$. The DPR then retrieves all documents (D$_1$,D$_2$,D$_3$,......,D$_n$) in the database along with their associated embeddings (E$_{D1}$, E$_{D2}$, E$_{D3}$,......,E$_{Dn}$). The embedded query E$_{question}$ is compared to each of the database-retrieved document embeddings. This comparison is carried out by computing the mathematical dot product and then taking the top k outcomes of these computations and producing their comparable relevant articles. At this point, we have the top k relevant articles from the database to the query. A copy of the question along with the top k relevant papers is fed into the reader model, which works on deriving the starting s and ending e locations of the response from the relevant articles. Lastly, it displays the first of the responses from these papers based on the starting and ending places.

The inputs as well as the outputs of such fine-tuned models are depicted in further detail in the following image. The Dense Passage Retriever DPR receives the query as an input, as shown in Figure 5. The DPR then encodes this question and attempts to find the most related articles for this question encoding. The dot-product of the encodings of all database articles and the question encoding is used to retrieve the most relevant articles. In this case study, the k value is 500. This retriever returns the top 500 articles linked to the question. This model obtains a Recall@500 of 87.8%.

As seen in Figure 6, our reader model takes the most similar articles as input together with a replica from the query. The reader selects the top 500 articles and outputs the response of the query.

This procedure involves retrieving the beginning and ending positions of the response from each article. Then, in the last column of Figure 6, it displays the responses ordered by the most relevant one concerning the score. This model has an F1 score of 75.8%.

7 Conclusion

Throughout this research, we explained the QA task in NLP and focused on the ODQA models, which were the intended task. In addition, recent studies related to this NLP task have been displayed. Following that, we demonstrated our ODQA system architecture, which included our EIAD dataset. This system architecture was based on a retriever-reader model, with two main models: reader and retriever. Each of these models had more than one trial. All of these trials were fine-tuned using the EIAD dataset. The most difficult challenge during this project was managing its resources. All trials were run on either a single RTX 2080ti GPU with 12GB VRAM or a single GTX 1080ti GPU with 12GB VRAM. Because DPR models contain two models internally as it is heavy, the RTX 2080ti was used to train them. The GTX 1080ti was used in conjunction with lighter models i.e., reader models. Another difficulty was the large size of the dataset. Our dataset (EIAD dataset) was 10k question-answer pairs in size, compared to the SQUAD dataset, which was 100k, and SUQAD2, which was 130k. Although we achieved new results for these models, the DPR model outperformed the reference model [7], which reached 78.8% top-100 accuracy with all MiniLM-L6-V2 and 78% recall@100 score while the reference model [7] reported 67.5% top-100 accuracy and 67% recall@100. Similarly, the reader model achieved 75.8% F1 score and 71.5% EM by using distilbert base-uncased-distilled-squad, which is the best outcome compared to [16], which reported 41.36% F1 score and 28.23% EM. As a result, we created a new baseline for the EIAD dataset called the all-MiniLM-L6-V2-base-uncased-distilled-squad-EIAD.

References

[1] A. Bouziane, D. Bouchiha, N. Doumi, and M. Malki, "Question Answering Systems: Survey and Trends," *Procedia Comput Sci*, 73:366–375, 2015, doi: https://doi.org/10.1016/j.procs.2015.12.005.

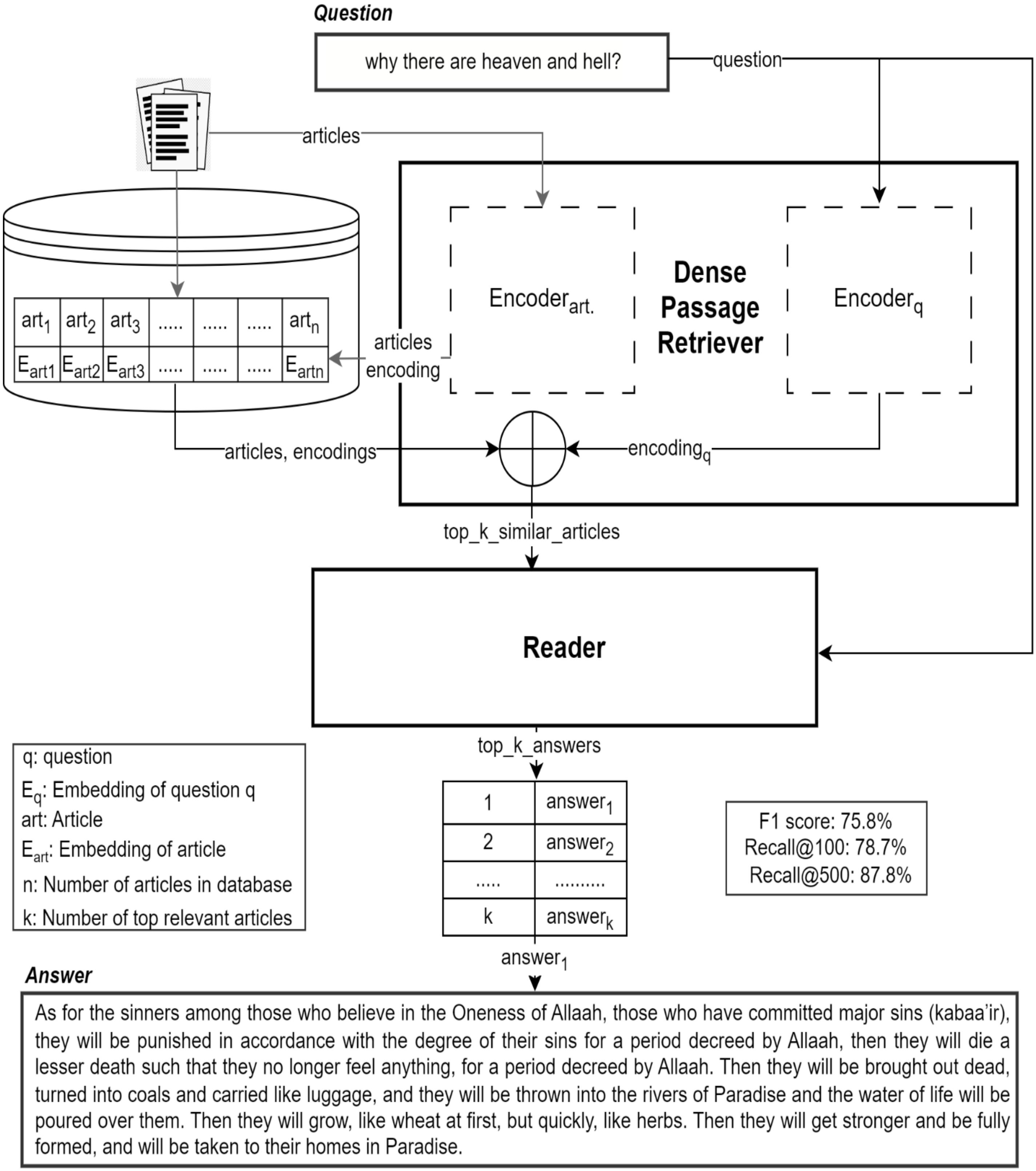
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *CoRR*, vol. abs/1810.04805, 2018.

[3] FAISS 2022, Accessed 30 December 2022, https://faiss.ai/.

[4] "Haystack Annotation Tool," https://www.deepset.ai/annotation-tool-for-labeling-datasets, Jan. 30, 2022.

[5] "Islam Question and Answer 2022," Access Jan. 25, 2022, https://islamqa.info/en.

[6] "Islam Religion 2022," Accessed Jan. 16, 2022, https://www.islamreligion.com/.

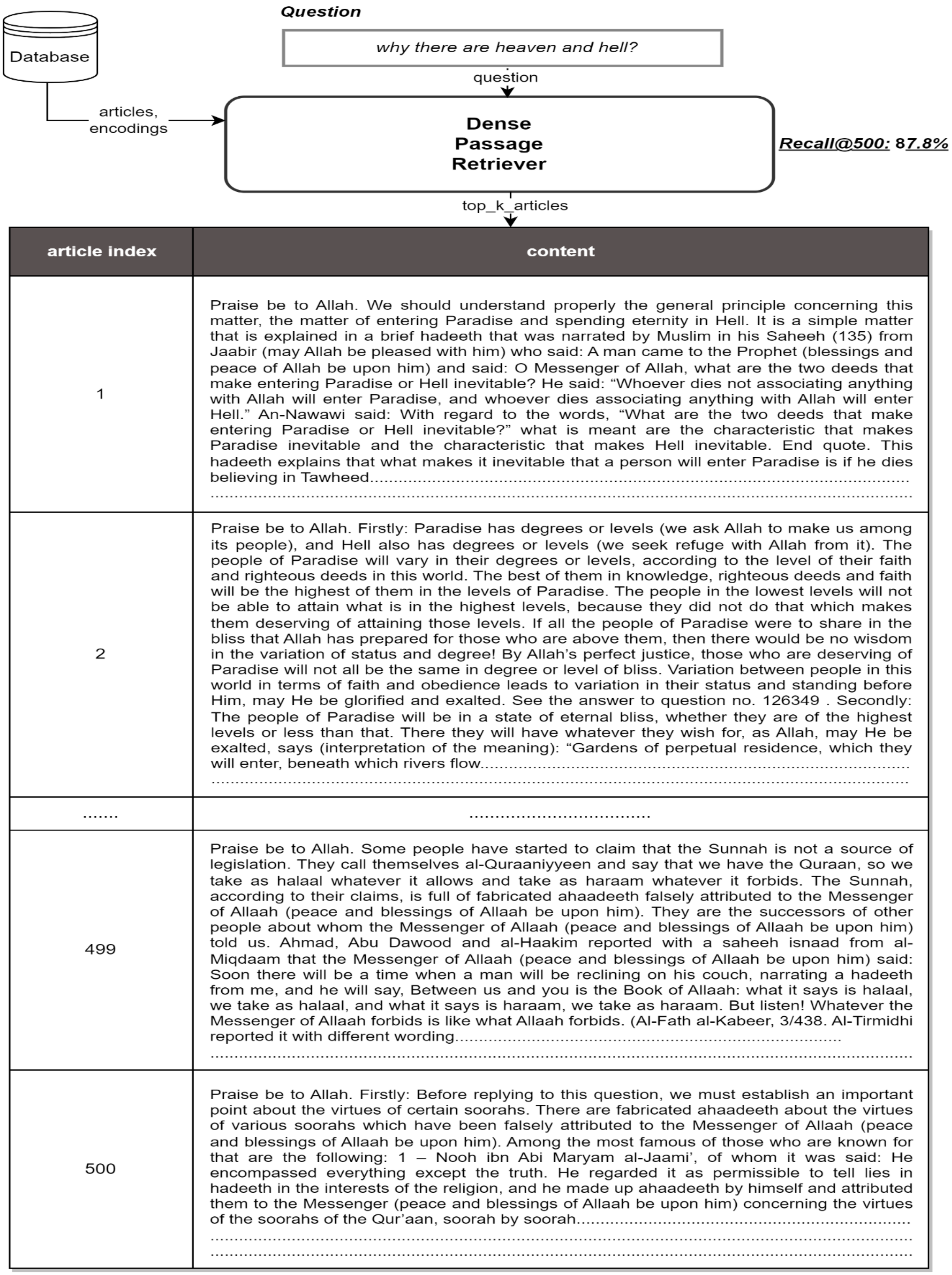


art$_1$	art$_2$	art$_3$	.....	.....	.....	art$_n$
E$_{art1}$	E$_{art2}$	E$_{art3}$	.....	.....	.....	E$_{artn}$

1	answer$_1$
2	answer$_2$
.....	....
k	answer$_k$

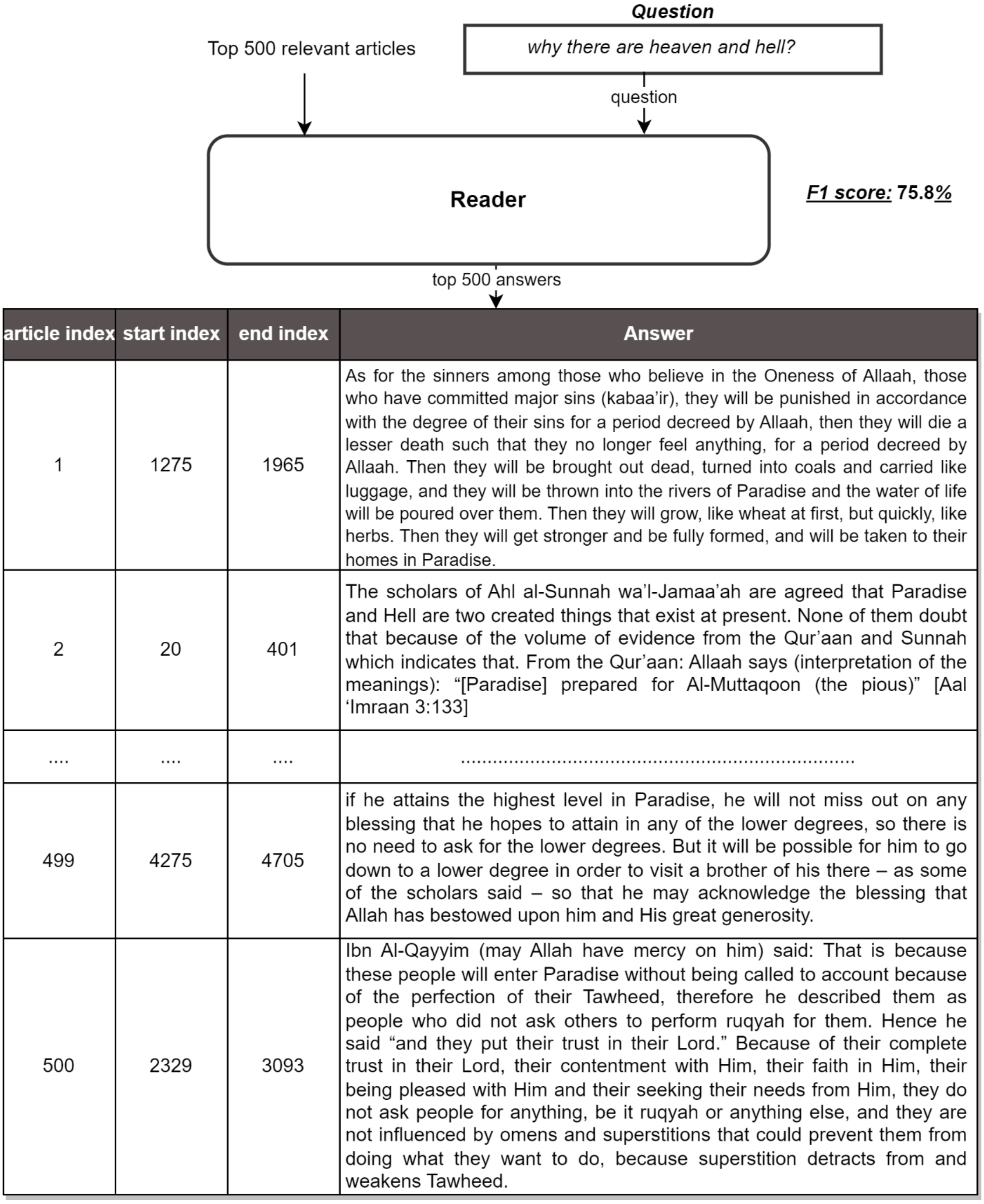
As for the sinners among those who believe in the Oneness of Allaah, those who have committed major sins (kabaa'ir), they will be punished in accordance with the degree of their sins for a period decreed by Allaah, then they will die a lesser death such that they no longer feel anything, for a period decreed by Allaah. Then they will be brought out dead, turned into coals and carried like luggage, and they will be thrown into the rivers of Paradise and the water of life will be poured over them. Then they will grow, like wheat at first, but quickly, like herbs. Then they will get stronger and be fully formed, and will be taken to their homes in Paradise.

Figure 4: Retriever reader case study



article index	content
1	Praise be to Allah. We should understand properly the general principle concerning this matter, the matter of entering Paradise and spending eternity in Hell. It is a simple matter that is explained in a brief hadeeth that was narrated by Muslim in his Saheeh (135) from Jaabir (may Allah be pleased with him) who said: A man came to the Prophet (blessings and peace of Allah be upon him) and said: O Messenger of Allah, what are the two deeds that make entering Paradise or Hell inevitable? He said: "Whoever dies not associating anything with Allah will enter Paradise, and whoever dies associating anything with Allah will enter Hell." An-Nawawi said: With regard to the words, "What are the two deeds that make entering Paradise or Hell inevitable?" what is meant are the characteristic that makes Paradise inevitable and the characteristic that makes Hell inevitable. End quote. This hadeeth explains that what makes it inevitable that a person will enter Paradise is if he dies believing in Tawheed.. ....
2	Praise be to Allah. Firstly: Paradise has degrees or levels (we ask Allah to make us among its people), and Hell also has degrees or levels (we seek refuge with Allah from it). The people of Paradise will vary in their degrees or levels, according to the level of their faith and righteous deeds in this world. The best of them in knowledge, righteous deeds and faith will be the highest of them in the levels of Paradise. The people in the lowest levels will not be able to attain what is in the highest levels, because they did not do that which makes them deserving of attaining those levels. If all the people of Paradise were to share in the bliss that Allah has prepared for those who are above them, then there would be no wisdom in the variation of status and degree! By Allah's perfect justice, those who are deserving of Paradise will not all be the same in degree or level of bliss. Variation between people in this world in terms of faith and obedience leads to variation in their status and standing before Him, may He be glorified and exalted. See the answer to question no. 126349 . Secondly: The people of Paradise will be in a state of eternal bliss, whether they are of the highest levels or less than that. There they will have whatever they wish for, as Allah, may He be exalted, says (interpretation of the meaning): "Gardens of perpetual residence, which they will enter, beneath which rivers flow.. ....
....	....
499	Praise be to Allah. Some people have started to claim that the Sunnah is not a source of legislation. They call themselves al-Quraaniyyeen and say that we have the Quraan, so we take as halaal whatever it allows and take as haraam whatever it forbids. The Sunnah, according to their claims, is full of fabricated ahaadeeth falsely attributed to the Messenger of Allaah (peace and blessings of Allaah be upon him). They are the successors of other people about whom the Messenger of Allaah (peace and blessings of Allaah be upon him) told us. Ahmad, Abu Dawood and al-Haakim reported with a saheeh isnaad from al-Miqdaam that the Messenger of Allaah (peace and blessings of Allaah be upon him) said: Soon there will be a time when a man will be reclining on his couch, narrating a hadeeth from me, and he will say, Between us and you is the Book of Allaah: what it says is halaal, we take as halaal, and what it says is haraam, we take as haraam. But listen! Whatever the Messenger of Allaah forbids is like what Allaah forbids. (Al-Fath al-Kabeer, 3/438. Al-Tirmidhi reported it with different wording.. ....
500	Praise be to Allah. Firstly: Before replying to this question, we must establish an important point about the virtues of certain soorahs. There are fabricated ahaadeeth about the virtues of various soorahs which have been falsely attributed to the Messenger of Allaah (peace and blessings of Allaah be upon him). Among the most famous of those who are known for that are the following: 1 – Nooh ibn Abi Maryam al-Jaami', of whom it was said: He encompassed everything except the truth. He regarded it as permissible to tell lies in hadeeth in the interests of the religion, and he made up ahaadeeth by himself and attributed them to the Messenger (peace and blessings of Allaah be upon him) concerning the virtues of the soorahs of the Qur'aan, soorah by soorah.. ....

Figure 5: Dense passage retriever top 500 documents



article index	start index	end index	Answer
1	1275	1965	As for the sinners among those who believe in the Oneness of Allaah, those who have committed major sins (kabaa'ir), they will be punished in accordance with the degree of their sins for a period decreed by Allaah, then they will die a lesser death such that they no longer feel anything, for a period decreed by Allaah. Then they will be brought out dead, turned into coals and carried like luggage, and they will be thrown into the rivers of Paradise and the water of life will be poured over them. Then they will grow, like wheat at first, but quickly, like herbs. Then they will get stronger and be fully formed, and will be taken to their homes in Paradise.
2	20	401	The scholars of Ahl al-Sunnah wa'l-Jamaa'ah are agreed that Paradise and Hell are two created things that exist at present. None of them doubt that because of the volume of evidence from the Qur'aan and Sunnah which indicates that. From the Qur'aan: Allaah says (interpretation of the meanings): "[Paradise] prepared for Al-Muttaqoon (the pious)" [Aal 'Imraan 3:133]
....	....	....	....
499	4275	4705	if he attains the highest level in Paradise, he will not miss out on any blessing that he hopes to attain in any of the lower degrees, so there is no need to ask for the lower degrees. But it will be possible for him to go down to a lower degree in order to visit a brother of his there – as some of the scholars said – so that he may acknowledge the blessing that Allah has bestowed upon him and His great generosity.
500	2329	3093	Ibn Al-Qayyim (may Allah have mercy on him) said: That is because these people will enter Paradise without being called to account because of the perfection of their Tawheed, therefore he described them as people who did not ask others to perform ruqyah for them. Hence he said "and they put their trust in their Lord." Because of their complete trust in their Lord, their contentment with Him, their faith in Him, their being pleased with Him and their seeking their needs from Him, they do not ask people for anything, be it ruqyah or anything else, and they are not influenced by omens and superstitions that could prevent them from doing what they want to do, because superstition detracts from and weakens Tawheed.

Figure 6: Reader extracting top 500 answers

[7] V. Karpukhin, B. Oguz, S. Min, L. Wu, S. Edunov, D. Chen, and W. Yih, “Dense Passage Retrieval for Open-Domain Question Answering,” *CoRR*, abs/2004.04906, 2020.

[8] K. Krishna, A. Roy, and M. Iyyer, “Hurdles to Progress in Long-form Question Answering,” *CoRR*, abs/2103.06332, 2021, Online: Available: https://arxiv.org/abs/2103.06332.

[9] J. Lee, M. Sung, J. Kang, and D. Chen, “Learning Dense Representations of Phrases at Scale,” *CoRR*, abs/2012.12624, 2020, Online: Available: https://arxiv.org/abs/2012.12624.

[10] S. Levy, K. Mo, W. Xiong, and W. Y. Wang, “Open-Domain Question-Answering for COVID-19 and Other Emergent Domains,” *CoRR*, abs/2110.06962, 2021, Online: Available: https://arxiv.org/abs/2110.06962.

[11] M. Mohammed, S. Amin, and M. M. Aref, “An English Islamic Articles Dataset (EIAD) for Developing an IslamBot Question Answering Chatbot,” *2022 5th International Conference on Computing and Informatics (ICCI)*, pp. 303–309, 2022, doi: 10.1109/ICCI54321.2022.9756122.

[12] “New Muslims 2022,” Accessed Jan. 15, 2022, https://www.newmuslims.com/.

[13] D. S. Sachan, M. Patwary, M. Shoeybi, N. Kant, W. Ping, W . L. Hamilton, and B. Catanzaro, “End-to-End Training of Neural Retrievers for Open-Domain Question Answering,” *CoRR*, abs/2101.00408, 2021, Online: Available: https://arxiv.org/abs/2101.00408.

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, J. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention is All You Need,” *CoRR*, abs/1706.03762, 2017, Online: Available: http://arxiv.org/abs/1706.03762.

[15] S. Wang, L. Zhou, Z. Gan, Y-C Chen, S. Sun, Y. Fang, Y. Chen, and J. Liu, “Cluster-Former: Clustering-based Sparse Transformer for Long-Range Dependency Encoding,” *ACL-IJCNLP 2021*, Aug. 2021.

[16] W. Yang, Y. Xie, A. Lin, X. Li, L. Tan, K. Xiong, M. Li, and J. Lin, “End-to-End Open-Domain Question Answering with BERTserini,” *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pp. 72–77, June 2019, doi: 10.18653/v1/N19-4013.

Moataz Mohammed is a teaching assistant of Computer Science Department at Ain Shams University, Cairo, Egypt. He is a Master candidate student in NLP field, B.Sc. of Computer Science Dept., Faculty of Computer and Information Sciences, in 2017, Ain Shams University, Cairo, Egypt.

Salsabil A. El-Regaily received her Ph.D. in 2019 and M.S. in 2013, both in computer science from University of Ain Shams, Cairo, Egypt. She also received a B.S. in 2007 in Computer Sciences from Ain Shams University, Cairo, Egypt. Her research interests are machine learning and medical image processing

Mostafa M. Aref is a Professor of Computer Science and Chairman of Computer Science Department, Ain Shams University, Cairo, Egypt. He received a Ph.D. of Engineering Science in System Theory and Engineering, June 1988, University of Toledo, Toledo, Ohio. He obtained his M.Sc. of Computer Science, October 1983, University of Saskatchewan, Saskatoon, Sask, Canada and a B.Sc. of Electrical Engineering - Computer and Automatic Control Section, in June 1979, Electrical Engineering Dept., Ain Shams University, Cairo, EGYPT.

UML Framework: National Institute for Health and Care Excellence (NICE) Diabetes Guidelines Based Diabetes Information System

Kalim Qureshi[*], Fatima Yousef[*], and Paul Manuel[*]
Kuwait University, Al-Shadadiya, KUWAIT

Abstract

Diabetes is a chronic disease that is a very common health problem around the world. Without proper care and treatment of diabetes, it could become a life-threatening disease. Currently, innovations in health information systems facilitate the patients and healthcare providers in the care and treatment processes. Diabetes management process is well-illustrated by the National Institute for Health and Care Excellence (NICE) diabetes guidelines. In this paper, we aim to propose a diabetes management information system (DMIS) based on NICE guidelines. The proposed DMIS is built on a design model through the Unified Modelling Language (UML). We validate the proposed DMIS and its UML models with respect to NICE recommendations by means of traceability techniques. The validation and verification were carried out involving software and health experts. The results show that our proposed DMIS model is comprehensive and incorporates all the NICE recommendations.

Key Words: Diabetes management information system, NICE guidelines, UML design, traceability techniques

1 Introduction

Diabetes Mellitus is one of the major chronic diseases in the world. It is a metabolic disorder that makes the body unable to produce insulin or incapable to efficiently use the produced insulin [19]. Due to the insulin deficiency, excessive amounts of blood glucose will be in the bloodstream. This may result in many health complications. Diabetes has two major types which are diabetes type 1 and type 2. In diabetes type 1, the body does not produce insulin, or produce insufficient amounts that do not serve the body's needs. The conventional treatment of diabetes type 1 requires insulin injections several times a day. On the other hand, in diabetes type 2, the body produces insulin but in an insufficient amount.

The initial stages of this type do not require insulin injections. Patients can successfully manage their condition by making careful adjustments to their lifestyles. If diabetes was not managed properly, it could lead to dangerous complications that can threat the patient's life. Hence, to avoid them and to live a healthy life, a proper management of the disease is required. Managing diabetes can be a tedious task due to the continuous attention it requires in monitoring the patients' conditions and following up with their medication. Various healthcare models to manage diabetes have been developed to overcome challenges associated with the management of diabetes [11, 18].

This paper proposes a Unified Modelling Language (UML) framework of a diabetes management information system (DMIS). The DMIS incorporates an intelligent decision support system that follows diabetes medical guidelines developed by the National Institute for Health and Care Excellence (NICE). After designing the model DMIS using UML, we carry out the validation and verification by the means of the traceability techniques from NICE recommendation to the system models of DMIS.

2 Literature Review

Information and Communication Technologies (ICT) play a key role in addressing health challenges [2]. A variety of e-health systems have been developed to provide a wide-range of healthcare services and share medical knowledge from different platforms by the utilization of ICT [7]. Examples of e-health applications include electronic health records, telemedicine, multimedia-based medical images, AI-based decision making and virtual reality based medical analysis that could be used by researchers, patients, healthcare providers, and health organizations [4]. Health stakeholders require information systems in order to store and process valuable medical data. In particular, patient require these information systems in order to monitor their health without the assistance of physicians [1].

Diabetes e-health systems and tools provide integrated platforms to serve the needs of the patients and healthcare providers [15]. Many health monitoring devices, sensors and mobile tools have been developed in the past decades and they have proven to enhance the quality of services provided to people who have long term or chronic conditions such as diabetes [10].

Management of diabetes requires tracking various aspects related to patients' conditions, including medication, physical activity, dietary intake and body reading. This would result in generating large amounts of data, and the new health related information technologies have provided user-friendly platforms for patients to manage their health conditions [17]. With the use of information technology and health information systems, it

[*] Department of Information Science. Emails: kalimuddinqureshi@gmail.com, fatimaash135@gmail.com, pauldmanuel@gmail.com.

became possible to provide prevention, diagnosis, treatment, and follow-up measures to patients in different settings [16]. Therefore, health authorities set high expectations in health informatics impacts on diabetes management [5].

Nowadays, a wide range of diabetes management systems have been developed to ease the lives of diabetic patients [5]. Reviewing these systems reveal that existing systems vary in different aspects such as the target population, restricted features, the medical instruments, and information system processes. In Table 1, we list some of the diabetes available solutions and discuss their main contributions. From the above literature survey, we have observed that different information systems have supported different features which are considered of central importance in the process of maintaining the disease effectively. These information systems have aimed to ease the life of diabetic patients by providing a platform that stores and operates diabetes related information. Some of these systems are moderately comprehensive as they have integrated a wide range of parameters and functionalities that help in achieving better control of diabetes and in coordinating the communication and data exchange between patients and healthcare providers. In Table 2, we provide a comparison between the features incorporated in the existing diabetes information systems. We list 15 features in Table 2.

Table 1: Literature description and key contributions

Reference	Description	Key Contributions
[9] Lelis & Motta [2018]	The paper presents a glucose management information system that supports patients in the management activities. The system uses a prediction rule-based method for glucose measurement based on glucose levels that have been collected previously.	▪ Monitor Glucose. ▪ Predict blood glucose levels. ▪ Collect food intake data. ▪ Collect data about insulin dosages.
[6] Gia et al. [2017]	The paper proposes a real-time continuous glucose monitoring system based on internet of things (IoT). Through the system healthcare providers and caregivers can easily monitor patients via web-browser or mobile application.	▪ Real time transmission of glucose and body temperature data. ▪ It uses NFR communication protocol to achieve energy efficiency.
[8] Kart et al. [2017]	The paper describes a clinical decision support and monitoring system for diabetes using evidence-based guidelines.	▪ Diagnose, and manage the treatment of diabetes based on medical guidelines. ▪ Monitor patient's glucose levels and other inputs. ▪ Generate reminders and motivational messages. ▪ Alert healthcare providers in critical situations.
[2] Adeyemo et al. [2016]	The paper discusses the development of an online support system for diabetes management which allows diabetic patients to provide detailed information about their health condition.	▪ Monitor different aspects of diabetes such as blood glucose, blood pressure, exercise, and food intakes. ▪ Medical reminders. ▪ Appointment booking. ▪ Real-time chat platform for healthcare providers and patients.
[17] Supriya & Rekha [2015]	Android based diabetes monitoring application (MediMinder) to monitor blood glucose levels and assist patients and healthcare providers to monitor the diabetes conditions.	▪ Sugar level monitoring. ▪ Medicine reminders. ▪ Decision support tool. ▪ Alert healthcare providers in critical situations. ▪ Generate medical reports.
[3] Akter & Uddin [2015]	The paper presents an Android application for diagnosis and treatment of diabetes and hypertension.	▪ Monitor blood glucose and blood pressure. ▪ Diagnose diabetes and hypertension. ▪ Generate medical reports.
[15] Sabbar & Al-Rodhaan [2013]	The paper discussed the implementation of a diabetes monitoring system on Android platform for managing and monitoring diabetic patients. The system monitors the daily data of diabetic patients in addition to arranging monthly visits to healthcare providers.	▪ Bluetooth enabled technology. ▪ Real-time data transmission of blood glucose levels. ▪ Monitor HbA1c levels. ▪ Remote communication with healthcare providers. ▪ System supports Arabic language.

Table 2: Features comparison of existing systems

Feature	[9] Lelis & Motta [2018]	[6] Gia et al. [2017]	[8] Kart et al. [2017]	[2] Adeyemo et al. [2016]	[17] Supriya & Rekha [2015]	[3] Akter & Uddin [2015]	[15] Sabbar & Al-Rodhaan [2013]
Diabetes diagnosis			✓			✓	
Monitor blood glucose	✓	✓	✓	✓	✓	✓	✓
Laboratory data (HbA1c…)			✓			✓	✓
Monitor blood pressure				✓		✓	
Monitor treatment (insulin, other medicines)	✓				✓		✓
Send medicine reminders				✓	✓		✓
Book appointments				✓			✓
Send appointment reminder				✓			✓
Communicate with healthcare providers				✓			✓
Decision support system (treatment, recommendation, monitoring plans)	✓		✓			✓	
Track physical activity			✓	✓			✓
Nutrition (Monitor food intakes, manage diet)	✓		✓	✓			✓
Alert healthcare providers in critical situations		✓			✓		
Integrate with other hospital departments							✓
Others (BMI, temperature, psychological state)		✓	✓			✓	
Generate medical reports					✓	✓	✓

3 Proposed SIBBIS System

We propose a diabetes management information system (DMIS) based on NICE diabetes guidelines. In 2015, the National Institute for Health and Care Excellence (NICE) produced well-defined guidelines that focus on Diabetes Mellitus [12, 13]. These guidelines are evidence-based and will help to develop a tailored individual therapy for each patient in achieving good control of diabetes. NICE guidelines target three key stakeholders:

1. Healthcare providers.
2. Diabetic patients and their families.
3. Hospitals and service institutions.

After a thorough study of NICE diabetes guidelines, we have gained an understanding of the features that must be considered when providing care and treatment for diabetic patients. NICE guidelines have provided detailed recommendations to successfully manage patients' health conditions and allow them to lead a healthy life. These recommendations cover the areas of care and treatment that must be managed to achieve a full control of patients' conditions. In Figure 1, we provide an abstract view of the proposed system. Figure 1 illustrates how the proposed DMIS is built systematically. Additionally, we want to add features extracted from NICE diabetes guidelines, and then check their traceability in the different system's models.

4 SIBBIS System Requirement

The proposed system's requirements have been extracted from NICE diabetes guidelines. The requirements of the proposed DMIS are:

- The system shall diagnose diabetes based on patient's medical information.
- The system shall provide education to patients about diabetes, self-management roles, and medical recommendations about lifestyle adjustments.
- The system shall manage blood glucose levels by reminding patients to measure it, identifying target values, and detecting risk values.
- The system shall manage patient's diet by tailoring nutritional advices and counting calorie intakes.
- The system shall manage patients' physical activities by providing tailored activity advices.
- The system shall manage patients' treatment by assessing

patient's medical date and recommending the appropriate treatment.

- The system shall manage blood pressure levels by reminding patients to measure it, identifying target values, and detecting risk values.
- The system shall manage the HbA1c levels, the test frequency and target level for each patient.
- Health complications associated with diabetes are managed by providing assessment to patients' symptoms and referring them to the appropriate hospital department.

In addition to these requirements, we aim to incorporate the requirements and features of existing diabetes information systems that have been reflected in Table 2. These requirements will be used by target DMIS stakeholders. The system's objects represent the stakeholders. These objects perform different activities in the system to fulfil their roles. Table 3 shows the system's main objects and the set of activities that each object ought to perform

To provide a high-level view of the system components, we have developed a context diagram of our proposed DMIS in Figure 2. The figure reflects the entities that interact with the system, their inputs to the system, and the outputs they receive from the system. The system will be used by multiple entities, each with a different set of roles. We emphasized the intelligent aspect of the system in red lines.

5 UML Design Models

We apply UML to describe our proposed DMIS requirements. UML is a very popular general-purpose software engineering

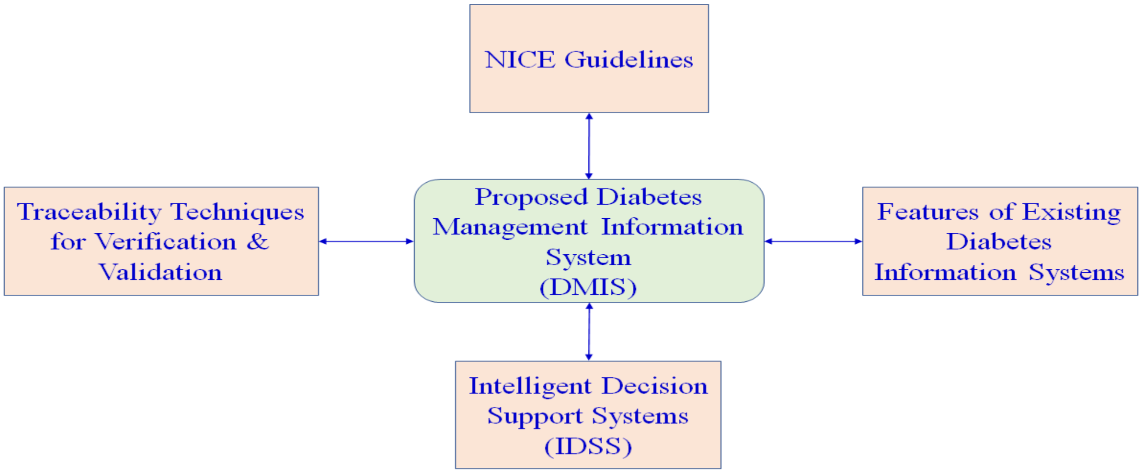

NICE Guidelines
Traceability Techniques for Verification & Validation
Proposed Diabetes Management Information System (DMIS)
Features of Existing Diabetes Information Systems
Intelligent Decision Support Systems (IDSS)

Figure 1: Abstract View of the proposed SIBBIS system

Table 3: SIBBIS system objects and stakeholders

OBJECT	ACTIVITY
Admin	create user account, authenticate user, update profile, manage accounts
Physician	diagnosis, view profile, request tests, view results, messages, referrals, medical recommendation, prescribe medicine
Lab Technician	conduct test, update results
Patient	make appointment, communicate to physicians, follow-up care and treatment, view results, evaluate doctor
Pharmacy	check medicine availability, update medicine records
Intelligent System	alerts, reminders, intelligent decisions, comprehensive medical recommendation
Dietician	view profile, update profile, determine diet plan and action plan
Physical Trainer	view profile, update profile, determine workout plan

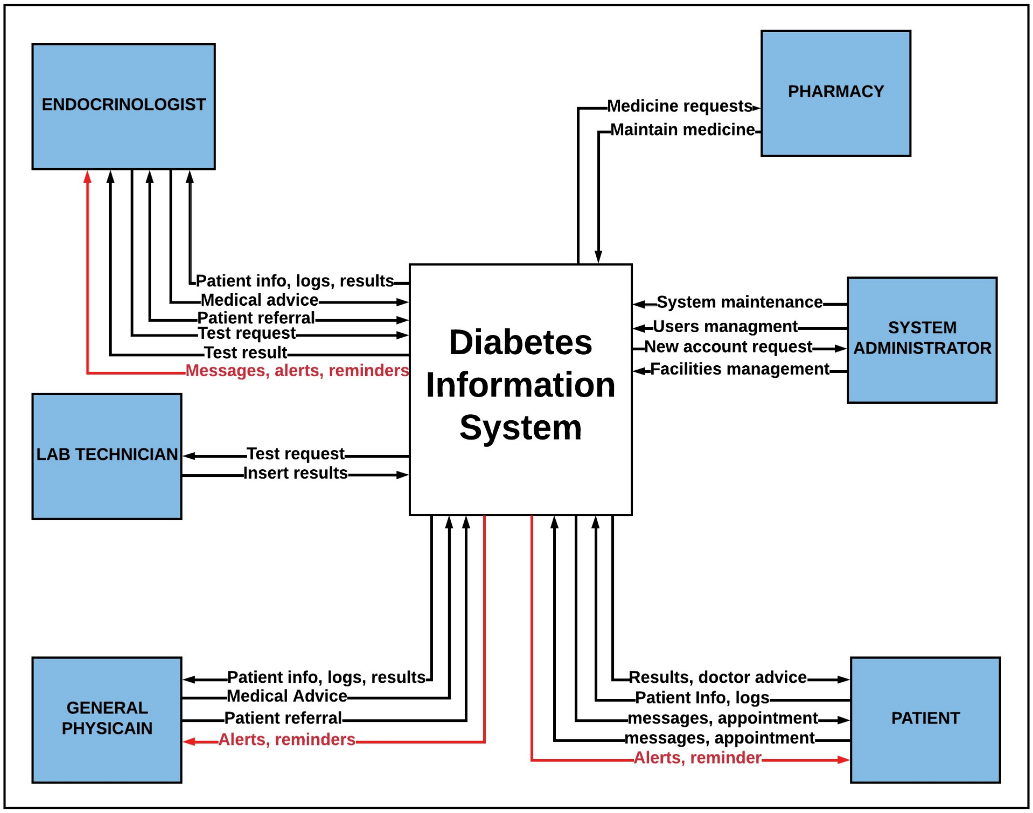


Figure 2: DIS context diagram

modelling language [14]. By utilizing it, one can develop a system's blueprint that assists in the following system's development and implementation phases. The features of UML will be used to guide software developers in specifying, visualizing, constructing, and documenting the artifacts of software systems. To present the diabetes system in a comprehensive manner, we develop different UML models that have different levels of complexity.

5.1 UML Class Diagram

The class diagram shows the overall structure of the system. Figure 3 shows the DMIS class diagram. In these diagrams, we can see that we have two base classes, "User" and "DiabetesCareAndTreatment", from which the rest of the classes are inherited. The first class identifies the system's users such as patient, system admin, and different healthcare providers. It shows the classes of these users interacting with the classes of the system and each of them has an account. The second class, DiabetesCareAndTreatment, reflects all the aspects that different users of the system interact with each other to deliver or receive services of diabetes care and treatment. This diagram will assist programmers in specifying the classes they must include in the implementation phase.

5.2 UML Use Case Diagram

The use case diagram is used to visualize the primary actors and their activities of the system. Through use case diagrams, we understand the functionalities of the system and the objects that initialized them. In Figure 4, we draw a high-level illustrative use case diagram of the DMIS. The diagram summarizes the relationships between the actors and the use cases. This diagram is used to show the primary requirements of the system without giving deep details or the exact order in which these functional requirements are performed.

The use case diagram in Figure 4 reflects the intelligent capabilities of the system. The intelligent decision support system is responsible for generating reminders related to appointments, medication, and medical check-ups. Moreover, the intelligent system will generate alarms at different levels in critical conditions. Also, it can produce different diabetes related management advice.

5.3Activity Diagram

Activity diagrams provide aggregated behavior model of the system. It shows the coordination of activities and how they flow when providing a service. Figure 6 reflects the diabetes diagnosis procedure and the users involved in this process. Several activities are conducted before reaching the final diagnosis. The intelligent system assesses the patient's health condition and suggests intelligent decisions to assist the physician. The physician views the recommendations and writes the final decision.

6 Traceability Checking

In this section, we check the traceability of NICE guideline recommendations in the different levels of the system's UML diagrams. NICE guideline processes have been reflected in diagrams that deal with specific aspects of the system's functionalities. The UML diagrams would permit the system developers to integrate the system's components. In the system development phase, the different UML diagrams with different levels of complexity would be combined. Therefore, establishing traceability between the different NICE recommendations in the UML diagrams is important.

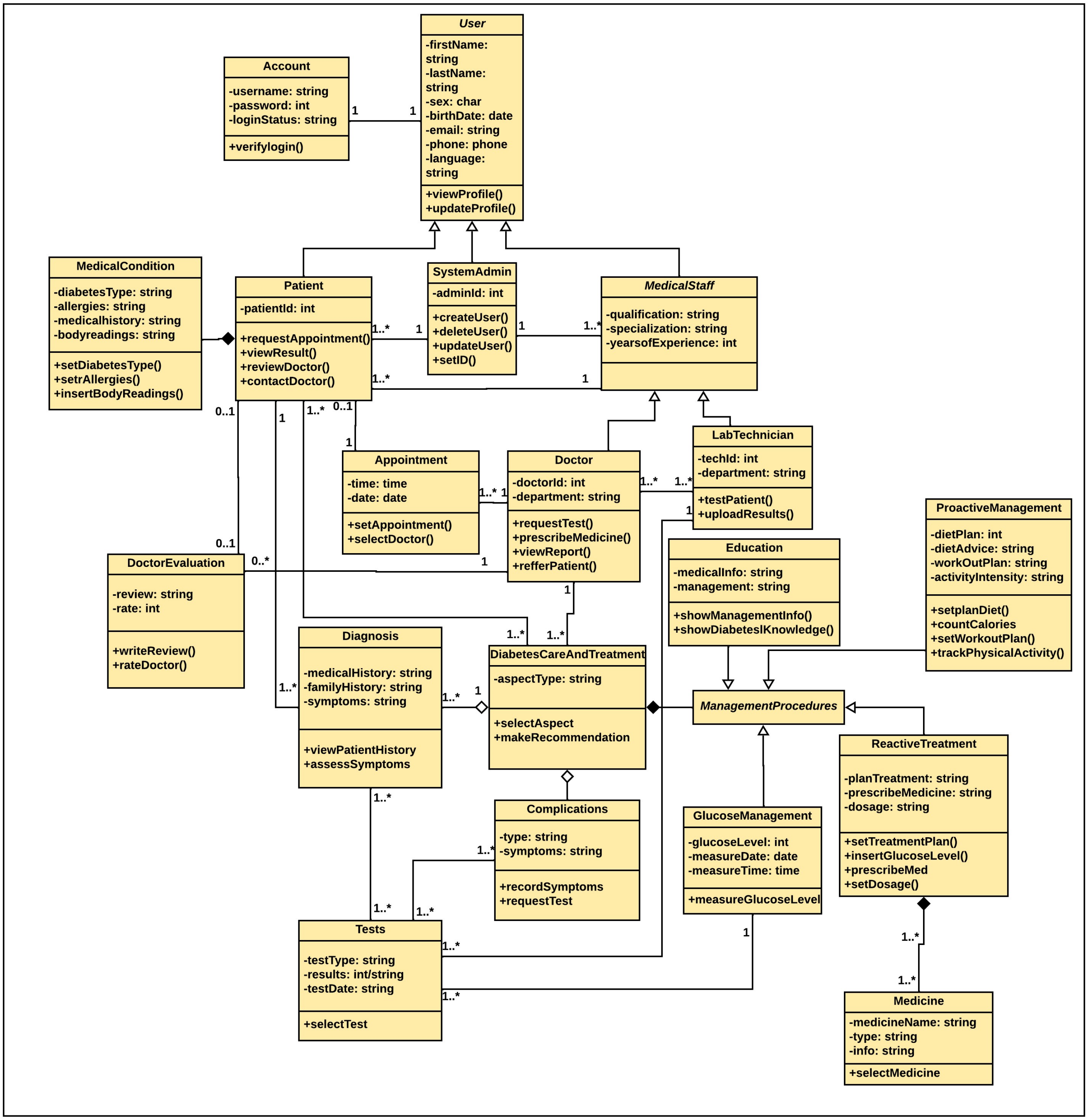


Figure 3: DMIS class diagram

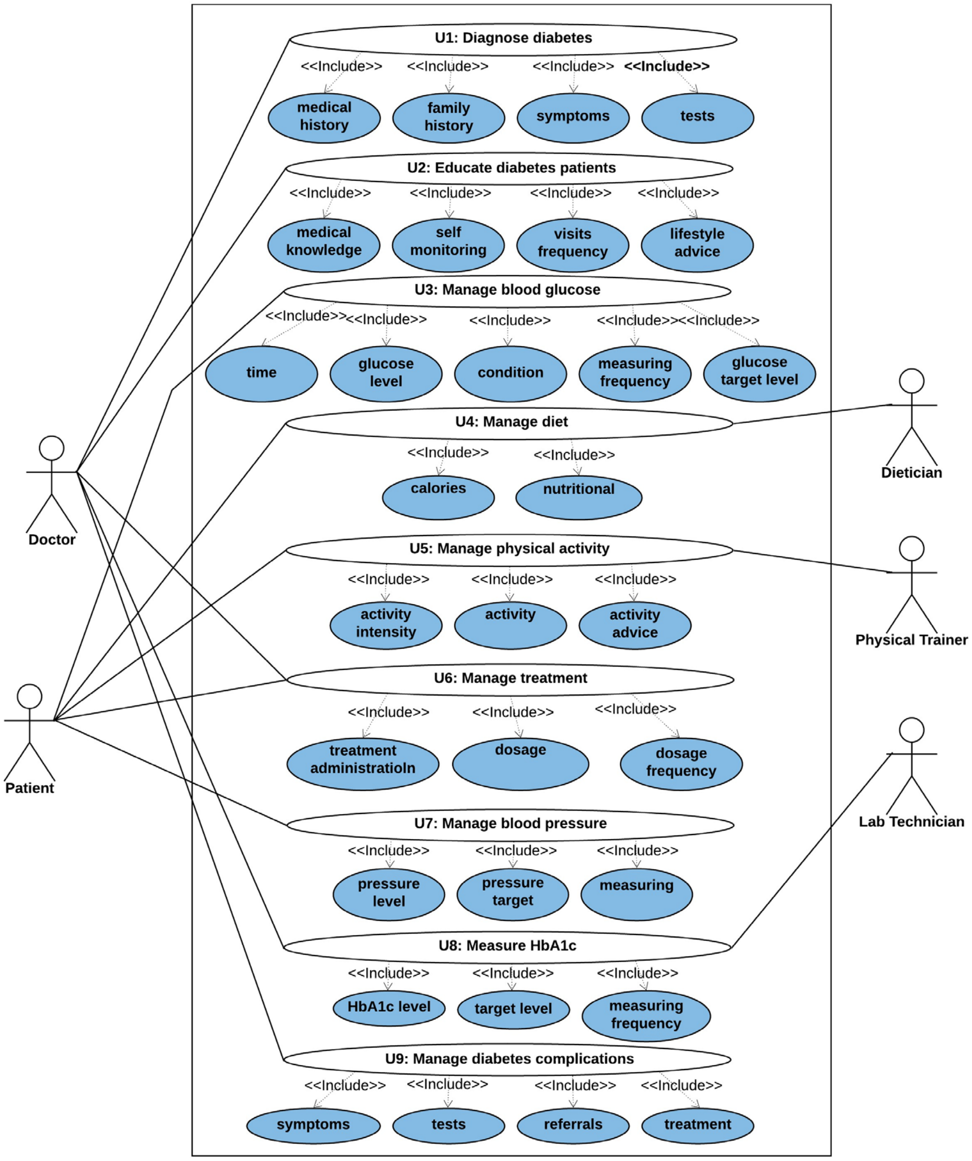


Figure 4: DMIS use case diagram

6.1 Traceability Checking Method

To check traceability, we have developed a questionnaire in which target participants were asked to check the traceability of NICE guideline recommendations in the different levels of the system's UML diagrams. The guideline recommendations have been listed to be initially checked against the system's use case diagram. Each recommendation has been represented in a stand-alone use case. Afterward, we continued checking the traceability of recommendations in corresponding activity diagrams and class diagram.

Instrument: A self-administrated questionnaire has been distributed to a few faculty members and IT graduate students. The first section of the questionnaire aimed to identify the respondents' profile. In the second section of the questionnaire, we have checked the traceability of NICE guideline diabetes recommendations in the UML diagrams.

Subjects: The questionnaire was administrated online using Google Forms. Our main target was collecting the responses of individuals who have software development knowledge. Therefore, the questionnaire link has been sent to faculty members and graduate students who took the system development and analysis course.

Questionnaire Design: We have developed a 5-point Likert scale. The scales were labelled as (1) VP= Very Poor; (2) P= Poor; (3) A= Acceptable; (4) G= Good; (5) VG= Very good. For the sake of efficiency, we have included the main system diagrams that reflect the integrated diabetes management

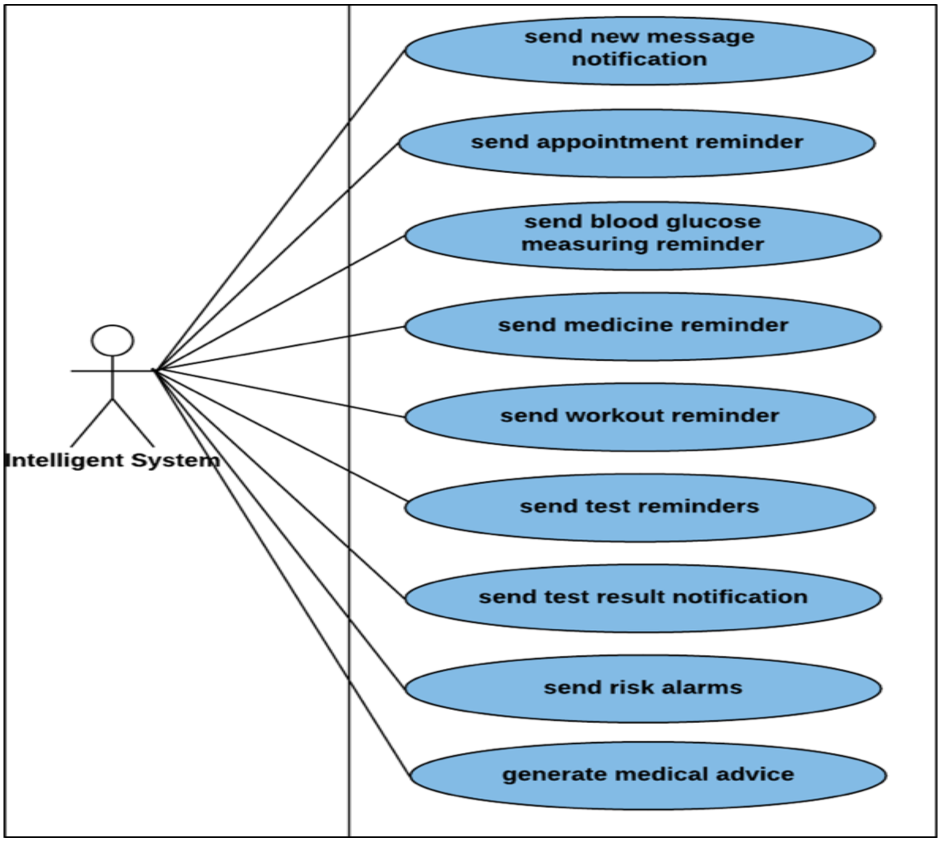

send new message notification
send appointment reminder
send blood glucose measuring reminder
send medicine reminder
send workout reminder
send test reminders
send test result notification
send risk alarms
generate medical advice
Intelligent System

Figure 5: Intelligent system use case diagram

processes. In the questionnaire, we have assessed 3 types of diagrams which are use case, activity, and design class diagrams. At first, users evaluate the traceability of the extracted NICE recommendations in the systems' use case diagram. Then, the traceability would be checked for each use case and its corresponding activity diagram. Finally, participants check the traceability of each activity diagram in the system's design class diagram. Respondents were requested to select one option per statement. From the responses, we would assess the quality of the UML design models for the extracted processes.

6.2 Measured Results

In this section we discuss the obtained results.

Demographics: Table 4, reflects the demographics of our participants. A number of 20 individuals have participated in our questionnaire. The age range of the participants was (20 and above 50). Most of the participants were in the age groups (21-30) and (31-40) years. In terms of gender, most of our participants were females. Moreover, participants were categorized based on their positions in the Information and Technology Department in Kuwait University into two groups, faculty member and graduate students.

Reliability Test: To validate data's reliability, we have used Cronbach's Alpha test in SPSS software. Alpha's value should be .7 or higher to be considered reliable. Table 5, shows the reliability results for the questionnaire's variables. The Alpha results show that all the variables have good reliability results.

Traceability between NICE Recommendations and their Corresponding Use Cases: In Table 6, we show descriptive statistics for the traceability of system's use cases and their source NICE diabetes recommendation. Through the results we would understand respondents point of view regarding the effectiveness of the system's use case diagram in reflecting the textual medical recommendation. The results show that a higher number of responses were on the good and very good sides. From that it is understood that our respondents agree that the use cases succeeded in reflecting nice recommendations in a comprehensive method. Thus, traceability between each use case and its source is established.

In Figure 7, we can see a graphical representation of the traceability between each use case and its corresponding NICE recommendatiuon.

Traceability between Use Cases and Their Corresponding Activity Diagrams: In Table 7, we display statistics for the traceability of activity diagrams to their corresponding use

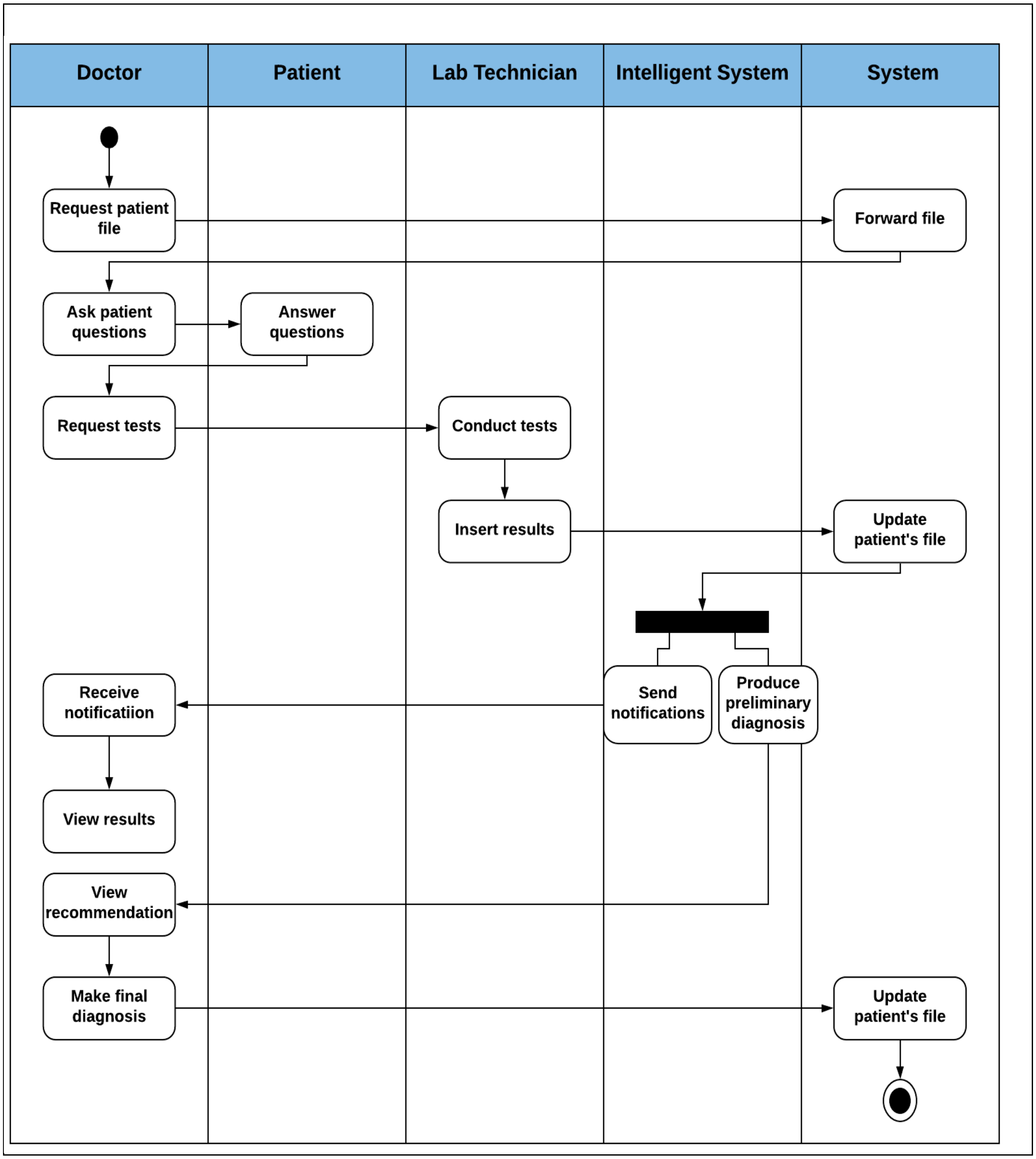


Figure 6: Diabetes diagnosis activity diagram

cases. From the results we can see if traceability of NICE recommendations is captured between the system's activity diagrams and use cases. Results show that the majority of responses agree that traceability between activity diagrams and their corresonding use cases is estaablished in a good and very good mnner. From this we understand that the activity diagrams were effective in reflecting NICE use cases.

Traceability between Activity Diagrams and their Corresponding Class in the Class Diagram: Table 8 shows the descriptive statistics of the traceability of design class diagram to the activity diagrams. From results we understand that the classes, in the class diagram, have captured the activity diagrams in good ways. Therefore, the components of the design class diagram can be traced back to more detailed diagrams.

In Figure 9, the graphical representation shows how the traceability between an activity diagram and its corresponding class in the class diagram was established.

Table 4: Demographics

Variable		**Frequencies**	**Percentage**
Gender	Male	7	35%
	Female	13	65%
Age Group	20-30	6	30%
	31-40	8	40%
	41-50	3	15%
	Above 50	3	15%
College Position	Faculty member	3	15%
	Graduate student	17	85%

Table 5: Reliability test

Variable	**Alpha**	**No. of Items**
Use Case Diagram	.909	9
Activity Diagrams	.976	8
Design Class Diagram	.938	9

Table 6: Traceability between NICE recommendations and use case diagram

Use case traceability to NICE recommendation	**U**	**VP**	**P**	**A**	**G**	**VG**
R1: Diagnose diabetes based on clinical grounds by considering symptoms and the history of a suspected patient.	**U1**	0%	0%	10%	70%	20%
R2: Offer structured education, to diabetes patients, that includes the following components: diabetes knowledge, management, self-monitoring, and lifestyle adjustments.	**U2**	0%	0%	20%	60%	20%
R3: Advise patients about glucose management and educate them about the target levels they must achieve different conditions, and tests frequency.	**U3**	0%	0%	15%	45%	35%
R4: Provide individuals with tailored dietary advice to control weight and manage blood glucose.	**U4**	0%	0%	25%	50%	25%
R5: Provide physical activity advice to diabetes patients and produce an activity program for those who choose to integrate physical activities into their lifestyle.	**U5**	0%	0%	20%	55%	25%
R6: Develop a treatment plan based on the diabetes type and the medical background of each patient.	**U6**	0%	0%	15%	55%	30%
R7: Advise diabetes patients about how to achieve the target blood pressure level, the measuring frequency and the procedures of avoiding its complications.	**U7**	0%	0%	10%	60%	30%
R8: Test HbA1c level in diabetes patients periodically, to assess their glucose management performance.	**U8**	0%	0%	15%	60%	25%
R9: Manage health complications associated with diabetes. Provide assessment to patients who have complication symptoms and refer them to the appropriate hospital department.	**U9**	0%	0%	15%	60%	25%

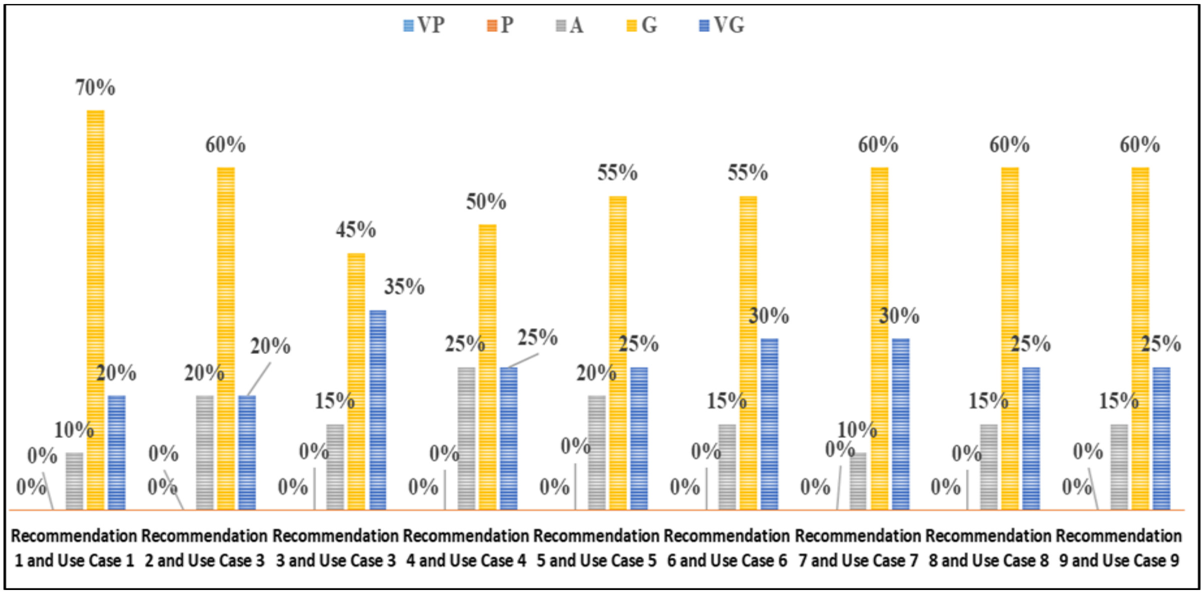


Figure 7: Traceability between use case diagram and NICE recommendations

Table 7: Traceability between activity diagrams

Activity Diagram Traceability to Use Case Diagram	VP	P	A	G	VG
Use Case 1 and Activity Diagram 1	0%	5%	20%	45%	30%
Use Case 2 and Activity Diagram 2	0%	0%	20%	50%	30%
Use Case 3 and Activity Diagram 3	0%	0%	30%	50%	20%
Use Case 4 and Activity Diagram 4	0%	5%	20%	50%	25%
Use Case 5 and Activity Diagram 5	0%	5%	25%	45%	25%
Use Case 6 and Activity Diagram 6	0%	0%	25%	45%	30%
Use Case 7 and Activity Diagram 7	0%	0%	20%	55%	25%
Use Case 8 and Activity Diagram 8	0%	0%	20%	55%	25%
Use Case 9 and Activity Diagram 6	0%	0%	25%	45%	30%

7 Discussion

In the previous sections we have checked traceability between NICE recommendations and the UML design models of our system. Our target respondents have checked the traceability between medical recommendations and the UML diagrams. Results have indicated that the traceability factor between the recommendation and the diagrams is achieved. Responses showed that the recommendations have been reflected in a good way in the use case diagrams. Traceability between the use case diagram and the corresponding activity diagrams and classes, in the design class diagram, was well established. Therefore, NICE recommendations were reflected very well in the system's UML designs.

8 Conclusion and Future Work

This paper discussed a diabetes information system DIS through the unified modelling language UML. The system's requirements have been extracted from diabetes medical guidelines developed by the National Institute for Health and Care Excellence (NICE). After identifying the requirements, we have modelled our system through UML. The system models aim to comprehensively reflect the functional requirements of the system that allow effective delivery of care and treatment services of diabetes. Additionally, we checked the traceability of NICE recommendation in the main system's UML models. We have distributed a questionnaire in the Information Technology Department in Kuwait University to faculty

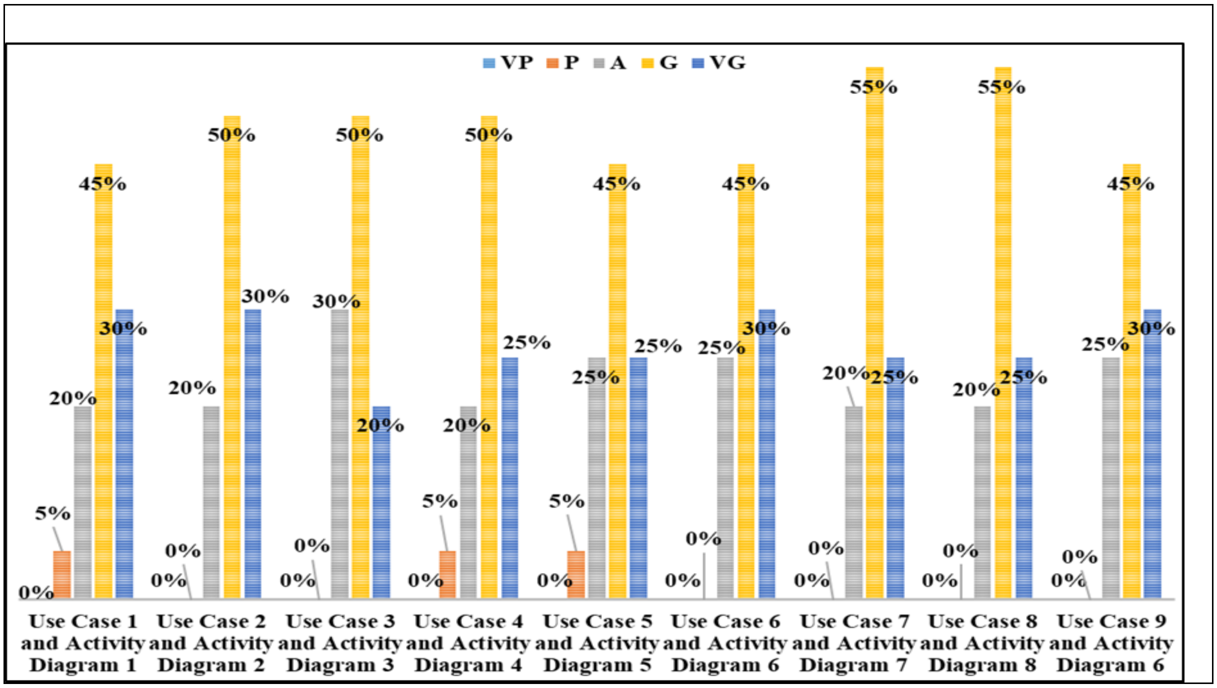


Figure 8: Traceability between activity diagrams and use cases

Table 8: Traceability between activity diagrams and class diagram

Activity Diagram Traceability to Class Diagram	VP	P	A	G	VG
Activity diagram 1 and Class 1	0%	0%	15%	55%	30%
Activity diagram 2 and Class 2	0%	0%	25%	50%	25%
Activity diagram 3 and Class 3	0%	0%	20%	55%	25%
Activity diagram 4 and Class 4	0%	0%	20%	50%	30%
Activity diagram 5 and Class 5	0%	0%	20%	45%	30%
Activity diagram 6 and Class 6	0%	0%	20%	50%	30%
Activity diagram 7 and Class 7	5%	10%	25%	40%	20%
Activity diagram 8 and Class 8	0%	0%	25%	45%	30%
Activity diagram 6 and Class 9	0%	0%	20%	20%	60%

members and graduate students. Results have shown that traceability of NICE recommendations in the systems' design models is achieved. With a traceability component we manage to reduce or avoid complications in the implementation phase. Additionally, we achieve consistency between the system's requirements and the design outcome. In the future, we plan to extend the work and incorporate security mechanisms. Additionally, we aim to develop a prototype of our system.

References

[1] Elizabeth Adejumo and Funmilola Ajala, "Health Monitoring System for Post-Stroke Management," *I. J. Information Engineering and Electronic Business*, 1:1-10, 2019.

[2] O A Adeyemo, O N Gidi-Fanimokun, and J O Alabi, "Online Support System for Diabetes Management," *International Journal of Computer Applications*, 152(10):6-11, October 2016.

[3] Morium Akterand and Mohammad Uddin, "Android-Based Diabetes Management System," *International Journal of Computer Applications*, 110:5-9, 2015, 10.5120/19350-0071.

[4] Abdelbaset M. Elghriani*, Abdelwanis A. Alabbar, Abdelsalam M. Maatuk, and Ehab A. Omar Elfallah, "Health Care Workers' Use of Electronic Medical

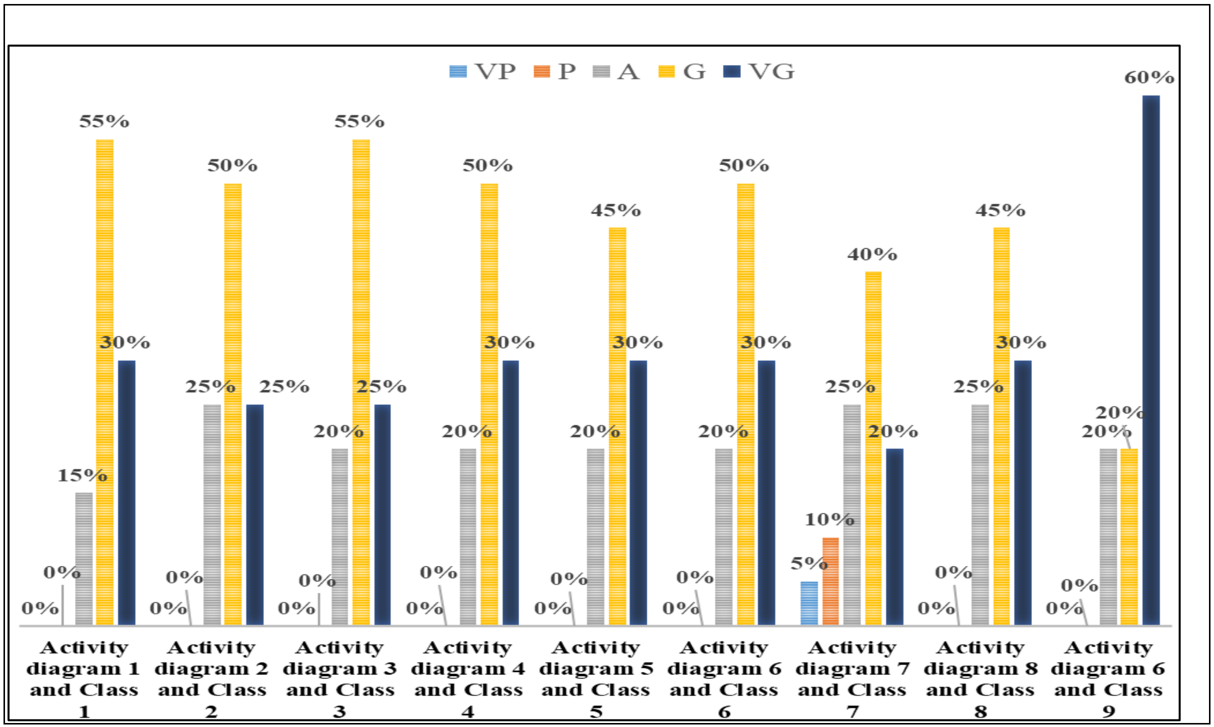


Figure 9: Traceability between activity diagrams and class diagram

Information Systems: Benefits and Challenges," International ACM Conference on Data Science, E-Learning and Information Systems, Ma'an, Jordan, 2021.

[5] Sylvia Franc, A. Daoudi, S. Mounier, B. Boucherie, D. Dardari, H. Laroye, B. Neraud, Elisabeth Requeda, L. Canipel, and Guillaume Charpentier, "Telemedicine and Diabetes: Achievements and Prospects," *Diabetes & Metabolism*, 37:463-76, 2011, 10.1016/j.diabet.2011.06.006.

[6] Tuan Nugyen Gia, Mai Ali; Imed Ben Dhaou, Amir M, Rahmani, Tomi Westerland, Pasi Liljeberg, and Hannu Tenhumen, "An IoT-Based Continuous Glucose Monitoring System: A Feasibility Study," Procedia Computer Science, 109. 10.1016/j.procs.2017.05.359, 2017.

[7] Conceiçao Granja, Wouter Janssen, and Monika Alise Johansen, "Factors Determining the Success and Failure of eHealth Interventions: Systematic Review of the Literature," *Journal of Medical Internet Research*; 20(5):e10235, 2018.

[8] Özge Kart, Vildan Mevsim, Alp Kut, İsmail Yürek, Oğuz Yılmaz, "A Mobile and Web-Based Clinical Decision Support and Monitoring System for Diabetes Mellitus Patients in Primary Care: A Study Protocol for a Randomized Controlled Trial," BMC Medical Informatics and Decision Making, 17, 10.1186/s12911-017-0558-6, 2017.

[9] Claudio Augusto Silveira Lelis and Renan Motta Goulart, "A Diabetes Management Information System with Glucose Prediction," *Journal of Information*, 9:319, 2018, doi: 10.3390/info9120.319.

[10] Soo Lim, Seon Kang, Hayley Shin, Hask Jong Lee, Ji Yoon, Sung Yu, Soi-Youn Kim, Soo Yoo, Hye Jung, Kyong Soo Park, Jun Ryu, and Hak Jang, "Improved Glycemic Control Without Hypoglycemia in Elderly Diabetic Patients Using the Ubiquitous Healthcare Service," *A New Medical Information System, Diabetes Care*, 34:308-13, 2011, 10.2337/dc10-1447.

[11] David Mulvaney, Bryan Woodward, S. Datta, Paul Harvey, Anoop Vyas, Bhaskar Thakker, Omar Farooq, and R. S. H. Istepanian, "Monitoring Heart Disease and Diabetes with Mobile Internet Communications," *International Journal of Telemedicine and Applications*, 195970. 10.1155/2012/195970, 2012.

[12] National Institute for Health and Care Excellence, "Type 1 Diabetes in Adults: Diagnosis and Management (NICE Guideline 17)," Available at: https://www.nice.org.uk/guidance/ng17, 2015, Accessed 13 May 2022.

[13] National Institute for Health and Care Excellence, "Type 2 Diabetes in Adults: Management (NICE Guideline 28)," Available at: https://www.nice.org.uk/guidance/ng28, 2015, Accessed 19 May 2022.

[14] James Rumbaugh, Ivar Jacobson, and Grady Booch, "The Unified Modelling Language Reference Manual," 2020.

[15] Mashael Sabbar and Mznah Al-Rodhaan, "Diabetes Monitoring System Using Mobile Computing Technologies," *International Journal of Computer Science and Applications*, 4:23-31, 2013, 10.14569/IJACSA.2013.040204.

[16] Remya Sivan and Zuriati Ahmad Zukarnain,"Security and Privacy in Cloud-Based E-Health System," *Journal of Symmetry*, 13(5):1-14, 2021.

[17] H. S. Supriya and Mrs. R. J. Rekha, "Medi Minder: A Blood Sugar Monitoring Application Using Android," *International Journal of Advanced Research in Computer and Communication Engineering*, 4(4):245-247, 2015 https ://www.ijarcce.com/upload/2015/april15/ IJARCCE% 2055.pdf.

[18] Xuejuan Wei, Hao Wu, Shuqi Cui, Caiying Ge, Li Wang, Hongyan Jia, and Wannian Liang, "Intelligent Internet-Based Information System Optimises Diabetes Mellitus Management in Communities," *Health Information Management Journal*, 47:70-76, 2017, 183335831769771. 10.1177/1833358317697717.

[19] World Health Organization, "Diabetes," Available at: https://www.who.int/news-room/fact-sheets/detail/ diabetes 2020, Accessed 20 April 2022.

Kalim Qureshi is an Associate Professor of Information Science Department, Kuwait University, Kuwait. His research interests include network parallel distributed computing, thread programming, concurrent algorithms designing, task scheduling, performance measurement and medical imaging. Dr. Qureshi received his Ph.D and MS degrees from Muroran Institute of Technology, Hokkaido, Japan in (2000, 1997). He published more than 70 journal papers in reputed journals. His email address: kalimuddin.qureshi@ku.edu.kw.

Fatima Yousef (photo not available) received a bachelor degree from English Linguistics and translation department of Kuwait University. She completed her master's in information technology from the department of Information Science, Kuwait University in 2020. Currently she is working as a real estate data analytic in the private sector.

Paul D. Manuel is a Professor in Information Science at Kuwait University. He received his M.S in Computer Science from the University of Saskatchewan, Canada in 1992. He obtained his Ph.D in Computer Science from the University of Newcastle, Australia in 1996. He got his first PhD in computing from Indian Institute of Technology, India in 1986.

His current research interests are in Information Systems, Graph Algorithms and Computational Complexity. According to Google Scholar, his publications have been cited more than 2567 times. The h-index is 28 and i10-index is 65. He has published more than over 100 scientific papers in peer reviewed journals with a total impact factor exceeding 70. He was listed in the 'World's Top 2% Scientist 2022' as one of the world's most influential scientists, which was released by Stanford University and Elsevier Foundation.

Building Computer-Based Test (CBT) using MATLAB: Programming the Essential Types of Questions

Baghdadi Ammar Awni Abbas* and Mohammed Al-Mukhtar†
University of Baghdad, Baghdad, IRAQ

Abstract

MATLAB is considered one of the most important multipurpose programs. In this paper we propose a testing package that can create Computer-Based Tests (CBT). The package contains the most frequent question types that are adopted in the most prominent Learning Management Systems (LMS) such as Google Classroom, Moodle, Canvas, Blackboard, D2L, Joomla, Schoology, and Talent. Six types of questions are discussed: True or False questions (TorF), Multiple Choice Questions (MCQ) single choice and multiple choice, fill in the blank, essay questions, and matching. Each of these questions is built using the MATLAB App Designer tool that comes with MATLAB R2020a. The package uses an Excel spreadsheet as a storage for the exam's information, student's answers, and grades. The user can make an exam with an unlimited number of questions. The user can take an exam with two options, either without any help as a real test, or taking the exam as training, where a button to "show the correct answers" is visible. The grading of the exam is a mixed operation between the user and the computer, the fill in the blank and essay question are graded manually. All other forms of questions are graded automatically. The graphical user interface is built for English language use. The exams are static, and no form of adaptation is used. Testing the program showed that the results are 100% accurate for a specimen of 200 users undertaking 20 different exams.

Keywords: Computer-based test, learning management system, MATLAB, true or false questions, multiple choice questions, fill in the blank, essay questions.

1 Introduction

There has been a gradual growth over the past 40 years. in CBT. as a suitable replacement for to paper and pencil testing. CBT was one of the most widely used internet uses in the late 1990s, but e-learning has recently gained significant importance, especially since the corona pandemic's emergence. Based upon reviewing the main LMS, six types of questions were selected to be built programmatically using the MATLAB App Designer.

The exam is a non-adaptive, fixed test where the question types and order are previously selected by the creator of the exam according to the type of materials examined. The user can navigate through the question using the next/previous button. The program has no time limit and the user must end the exam manually to get the final score and get the certificate to pass the exam. The program presented runs on a standalone computer. The user creates an exam and automatically the exam is stored in an Excel spreadsheet named after the exam name selected by the creator. The spreadsheet in the Excel file serves as an exam bank, where the creator sets a complete exam in every spreadsheet.

2 Related Work

In [3] a comparison between CBT and paper-based exams for the postgraduate student is made. The majority of students preferred the first over the latter. Grading MCQ automatically without human interaction in [4], the program has 100% results accuracy. In [8] a fingerprint method is used to authenticate the examinee; the level of authentication is highly improved using these techniques. In [15] a biometric recognition method is used to authenticate exam entrance. In [16], methods and models of creating CBT are thoroughly explained, describing the ways to compare the models and describing the test delivery methods, finally; the validity issues are discussed; showing that it's a key issue when deciding the best model for the program. An application for CBT for smartphones (android) is presented in [14], the satisfaction for a sample of 30 students and a teacher is measured through a questionnaire, and a high rate of acceptance is noted. A novel approach to CBT is introduced in [12], where the examinee is asked to assemble objects on a computer screen; the test is made using Macro Media Flash. A comparison between traditional MCQ types of questions and innovative item formats in a CBT program was analyzed for IRT information with the three parameter and graded response models [9]. The waterfall model and the Reuse-oriented software process models are used in [2] to make a component-based software that recycles the same software element to make other components. A way to evaluate and minimize the length of a CBT on sentence comprehension is presented in [19], 5 to 8 minutes were reduced from the exam time with the same results. The analysis in [18] discusses the effectiveness of the bimasoft application as a medium for evaluating CBT learning using Android. Results of the study found that the application of the

* College of Mass Media. Email: ammar.a@comc.uobaghdad.edu.iq
† Computer Center.

bimasoft application as a medium for evaluating CBT learning using Android was very good (84.64%). In [6] light is shed on the increase of electronic tools in education and learning, focusing on terminology, with diverse terms utilizing the same assessment approach within the literature, such as electronic assessment/evaluation, and online assessment/evaluation. The work [13] is the closest to our work, where a system to create/undertake exams is presented; the system is built using visual basic, HTML, and My SQL. The system proved to be very robust, stable, and error-free. The results of favoring computer-based testing complex "Profvybir" are presented in [10], 780 students took part in the assessment. This paper's method of occupational guidance can be used to implement the policy of the Ministry of Education and Science in the development of occupational guidance can be applied by a career advisor at school, while the CBT complex may be an additional component of the occupational training program. In [5] solutions to the current problems of CBT in Nigeria presented the user-friendly program using visual basic. The system is designed using the agile methodology through the extreme programming approach and unified modeling language was used to bring the view to real-life situations. Software is meant to be used for all kinds of CBT conducted or managed by universities in Nigeria. The top-to-down approach was adopted as the implementation approach for the development. The study of accessibility problems in CBT for blind persons is studied in [17] and gives some recommendations to facilitate the ease of use for them, the results showed that most CBT does not meet the expectations of visually impaired persons. In [1] there is discussion of the use of blended learning in Iraqi universities, where the proposed hybrid model has decreased the error rate from (0.00014) to (0.00013). An overview of the CBT models is made in [11]. Nigerian undergraduate students, participate in a CBT by comparing several modules studied by the students, the results of [7] showed that the students prefer CBT over paper-based exams.

2.1 Software Description

The main graphical user interface Figure 1 starts with a simple window that takes the user to three main parts of the program: 1-creating the exams 2-taking an exam that is made in the first part, either as a real exam or as a practice (which shows the correct answer button hidden according to the type of exam) 3-grading an exam taken in part two of the program. The fourth part of the main interface contains a brief description of the program and the programmer.

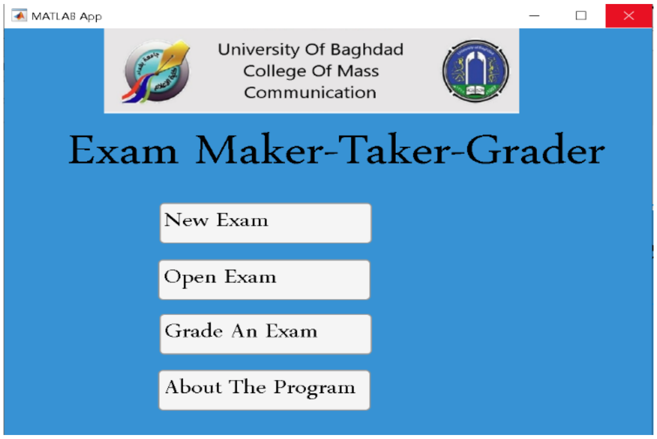


Figure1: Main graphical user interface

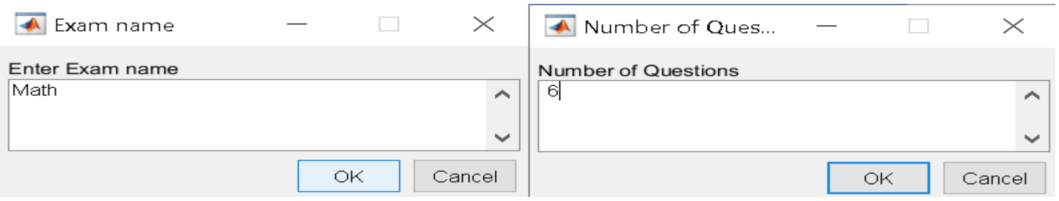
2.2 Part One: Creating a New Exam

The first part of the main GUI deals with creating a new exam, when this button is pressed, a message box popup asks the user to enter a proper name for the exam followed by another message box to specify the number of questions in the exam, as shown in Figure 2. The coding for these message boxes is illustrated in Code 1.

The first message will be turned into a spreadsheet in an Excel file that contains all the necessary information about the exam and the correct answers (including the total number of questions obtained from the second message box). Afterward, a dropdown menu appears asking the user to choose one of the six available types of questions as shown in Figure 3. The first row of the spreadsheet is reserved for the name of the exam cell (1,1) and cell (1,2), and the second row is also reserved for the number of the question in cells (2,1) and (2,2). Code 2 presents the line way to choose from the drop-down box.

There are six types of questions in the program:

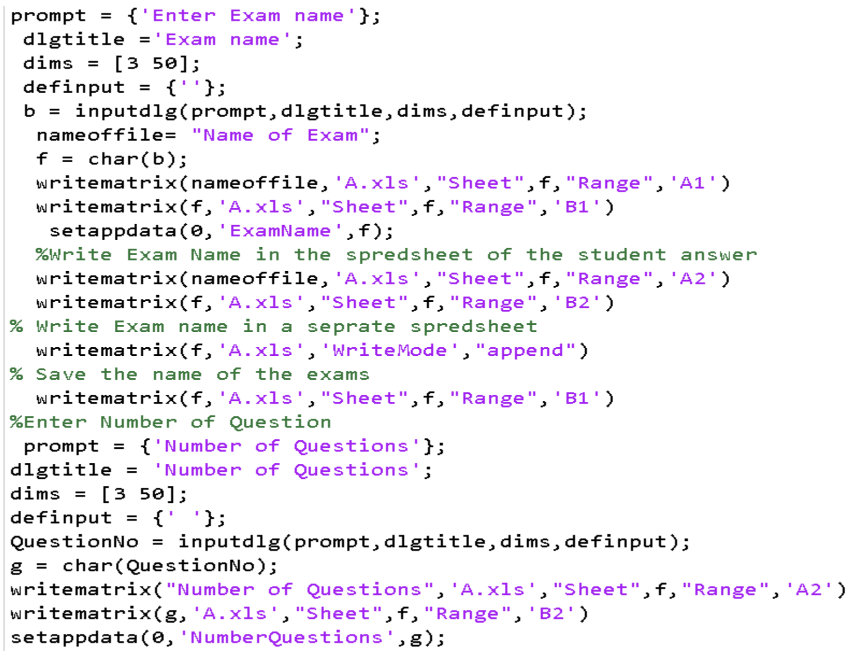
1-True or False: The coding for this type of question in the Excel spreadsheet is made as followed: The first cell (x,1) is the code of the question which is given the number 1, and to the right a cell that contains the question cell(x,2) followed by a third cell for the correct answer (1 denotes a True answer and 2 denotes a False answer). The GUI for T or F question is shown in Figure 4. Code 3 presents the programming of the same question after retrieving the correct name of the spreadsheet, the total number of questions, and the current question pointer position.



Enter Exam name
Math

Number of Questions
6

Figure 2: Getting the name of the new exam with the total amount of questions



```
prompt = {'Enter Exam name'};
 dlgtitle ='Exam name';
 dims = [3 50];
 definput = {''};
 b = inputdlg(prompt,dlgtitle,dims,definput);
  nameoffile= "Name of Exam";
  f = char(b);
  writematrix(nameoffile,'A.xls',"Sheet",f,"Range",'A1')
  writematrix(f,'A.xls',"Sheet",f,"Range",'B1')
   setappdata(0,'ExamName',f);
  %Write Exam Name in the spredsheet of the student answer
  writematrix(nameoffile,'A.xls',"Sheet",f,"Range",'A2')
  writematrix(f,'A.xls',"Sheet",f,"Range",'B2')
% Write Exam name in a seprate spredsheet
  writematrix(f,'A.xls','WriteMode',"append")
% Save the name of the exams
  writematrix(f,'A.xls',"Sheet",f,"Range",'B1')
%Enter Number of Question
 prompt = {'Number of Questions'};
dlgtitle = 'Number of Questions';
dims = [3 50];
definput = {'  '};
QuestionNo = inputdlg(prompt,dlgtitle,dims,definput);
g = char(QuestionNo);
writematrix("Number of Questions",'A.xls',"Sheet",f,"Range",'A2')
writematrix(g,'A.xls',"Sheet",f,"Range",'B2')
setappdata(0,'NumberQuestions',g);
```

Code 1 Coding the message box and changing the input to a spreadsheet

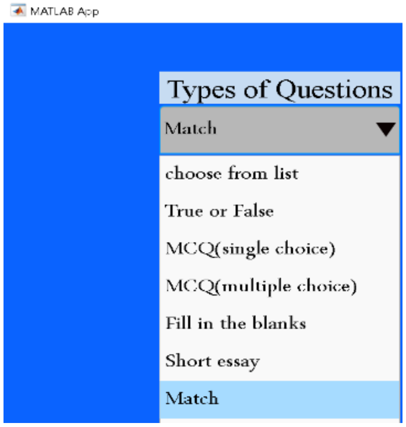
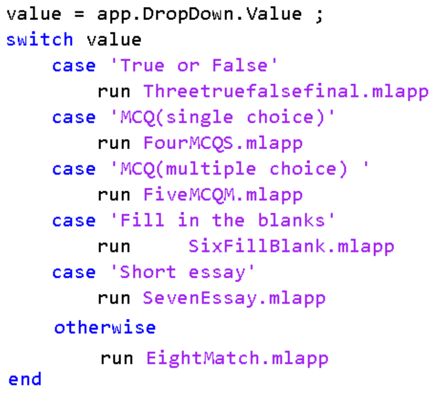


Figure 3: Question types



```
value = app.DropDown.Value ;
switch value
    case 'True or False'
        run Threetruefalsefinal.mlapp
    case 'MCQ(single choice)'
        run FourMCQS.mlapp
    case 'MCQ(multiple choice) '
        run FiveMCQM.mlapp
    case 'Fill in the blanks'
        run    SixFillBlank.mlapp
    case 'Short essay'
        run SevenEssay.mlapp
    otherwise
        run EightMatch.mlapp
end
```

Code 2: Choosing from drop down menu items

After typing the question and choosing the correct answer the creator must push the save button and wait for the confirmation message box “Your Question is Saved” before choosing to return to the main page to pick another form of question or staying with the same type of question.

2-Multiple Choice Questions (a single Choice from four choices): The coding for this type of question in the Excel spreadsheet is made as follows: The first cell (x,1) is the code of the question which is given the number 2, to the right a cell that contains the question cell(x,2) followed by a third cell (x,3) for the correct answer (1 denotes a first answer is correct, 2 denotes a second answer is correct, 3 denotes a second answer is correct and 4 denotes the fourth answer is the right answer). The cells(x,4) to (x,7) contain the four choices for the question. This type of question is made using a radio button which enables only one choice for each question. The GUI for MCQ single-choice question is shown in Figure 5, while the coding for such a question is presented in Code 4.

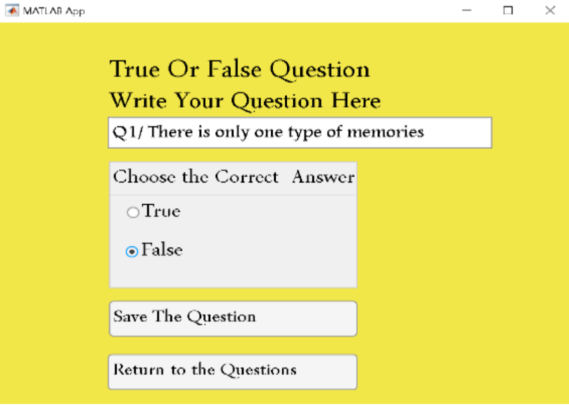
3-Multiple Choice Questions (with multiple choices): This type of question uses checkboxes instead of radio buttons which allows the user to select more than one option. The coding for this type of question in the Excel spreadsheet is made as follows: The first cell (x,1) is the code of the question which is given the number 3, to the right is a cell that contains the question cell(x,2) followed by a third cell (x,3) for the correct answer (see Table 1 for the different combinations of the answer). The cells(x,4) to (x,7) contain the four choices for the question. The MCQ multiple choices question is shown in Figure 6.

The MCQ (multiple choice) has almost the same GUI as the MCQ (single choice) with the same save and back to questions GUI buttons. The coding for that type of question is presented in Code 5.

4-Fill in the blank: The coding for this type of question in the Excel spreadsheet is made as followed: The first cell (x,1) is the code of the question which is given the number 4, and to the right is a cell that contains the question cell(x,2) followed by cell (x,3) to (x,6) for the correct answers. This type of question is graded manually unlike the first three types of questions which are graded automatically by the program. The fill in the blank question is shown in Figure 7. The coding for fill in the blank is presented in Code 6.

5-Essay Questions: The student must write the answer in the edit text field. The coding for this type of question in the Excel spreadsheet is made as followed: The first cell (x,1) is the code of the question which is given the number 5, to the right is a cell that contains the question cell(x,2) followed by cell (x,3) for the correct answers. This type of question is also graded manually like the fill in the blank question. Essay GUI is shown in Figure 8. The coding for the essay question is presented in Code 7.

6-Match Questions: The student is asked to match the options on the right to the ones on the left; the matching is made by clicking the number that carries the first sentence on the right first then clicking the matching sentence on the left. The user is asked to enter the correct option on the left that match the sentence on the right and this was considered the correct answer. In a real exam both the left and right columns of options are distributed randomly each time, an exam is opened (as a real exam or as a training exam). A click on the number of the sentence will change the color of the number and if it was followed by clicking the matching option on the left, the same color will appear on the left. The coding for this type of question in the Excel



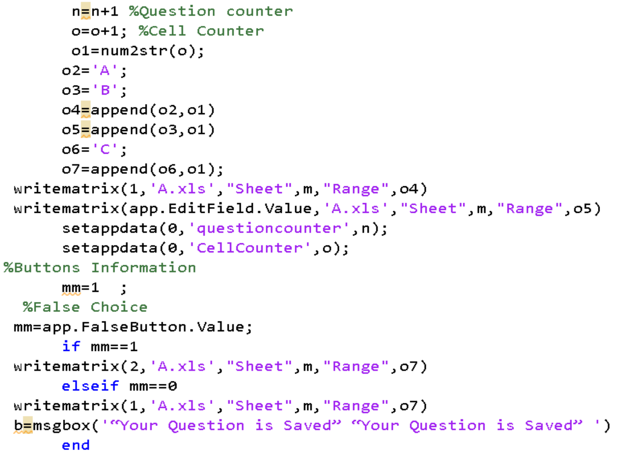
True Or False Question
Write Your Question Here

Q1/ There is only one type of memories

Choose the Correct Answer

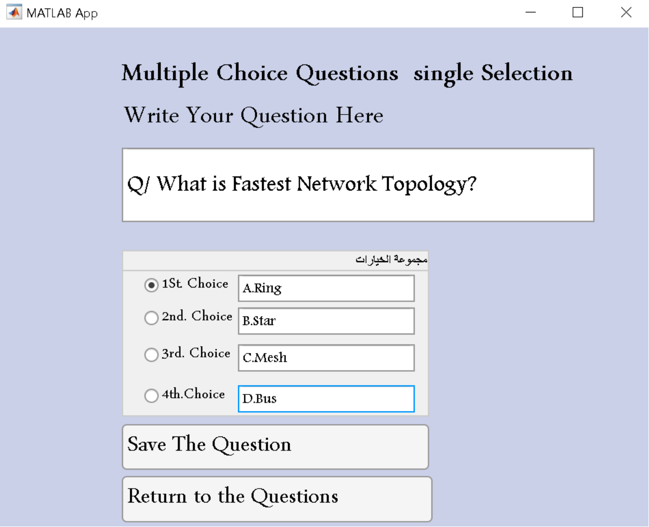
○ True
● False

Figure 4: True or false question type



```
       n=n+1 %Question counter
       o=o+1; %Cell Counter
       o1=num2str(o);
      o2='A';
      o3='B';
      o4=append(o2,o1)
      o5=append(o3,o1)
      o6='C';
      o7=append(o6,o1);
 writematrix(1,'A.xls',"Sheet",m,"Range",o4)
 writematrix(app.EditField.Value,'A.xls',"Sheet",m,"Range",o5)
      setappdata(0,'questioncounter',n);
      setappdata(0,'CellCounter',o);
%Buttons Information
      mm=1  ;
  %False Choice
 mm=app.FalseButton.Value;
      if mm==1
 writematrix(2,'A.xls',"Sheet",m,"Range",o7)
      elseif mm==0
 writematrix(1,'A.xls',"Sheet",m,"Range",o7)
 b=msgbox('“Your Question is Saved” “Your Question is Saved” ')
      end
```

Code 3: The T or F programming



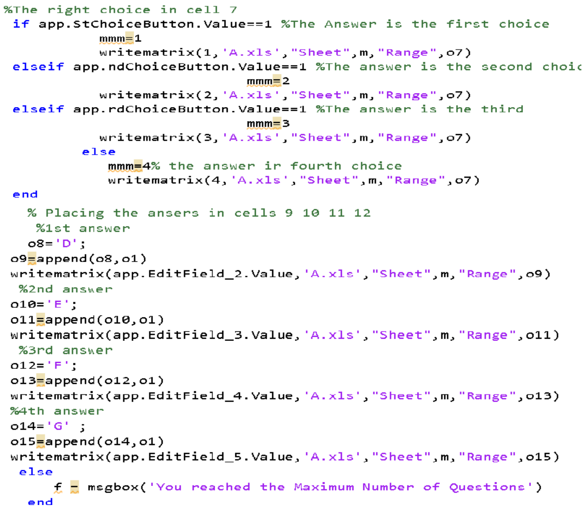
Multiple Choice Questions single Selection
Write Your Question Here

Q/ What is Fastest Network Topology?

مجموعة الخيارات

● 1St. Choice A.Ring
○ 2nd. Choice B.Star
○ 3rd. Choice C.Mesh
○ 4th.Choice D.Bus

Figure 5: Multiple choice questions (a single choice from four choices)

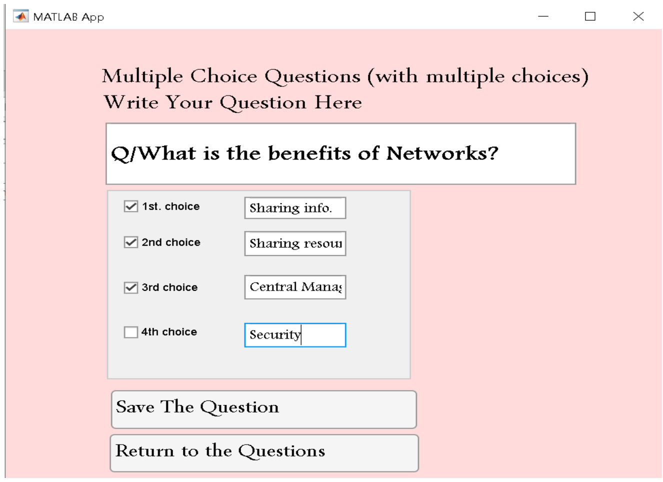


```
%The right choice in cell 7
 if app.StChoiceButton.Value==1 %The Answer is the first choice
             mmm=1
             writematrix(1,'A.xls',"Sheet",m,"Range",o7)
 elseif app.ndChoiceButton.Value==1 %The answer is the second choic
                                 mmm=2
             writematrix(2,'A.xls',"Sheet",m,"Range",o7)
 elseif app.rdChoiceButton.Value==1 %The answer is the third
                                 mmm=3
             writematrix(3,'A.xls',"Sheet",m,"Range",o7)
           else
              mmm=4% the answer ir fourth choice
              writematrix(4,'A.xls',"Sheet",m,"Range",o7)
 end
   % Placing the ansers in cells 9 10 11 12
    %1st answer
   o8='D';
o9=append(o8,o1)
writematrix(app.EditField_2.Value,'A.xls',"Sheet",m,"Range",o9)
  %2nd answer
o10='E';
o11=append(o10,o1)
writematrix(app.EditField_3.Value,'A.xls',"Sheet",m,"Range",o11)
  %3rd answer
o12='F';
o13=append(o12,o1)
writematrix(app.EditField_4.Value,'A.xls',"Sheet",m,"Range",o13)
%4th answer
o14='G' ;
o15=append(o14,o1)
writematrix(app.EditField_5.Value,'A.xls',"Sheet",m,"Range",o15)
  else
       f = msgbox('You reached the Maximum Number of Questions')
   end
```

Code 4: Programming MCQS

Table 1: The different possible combinations of MCQ (multiple choice) answer

Case No.	CheckBox1	CheckBox2	CheckBox3	CheckBox4	Answer No.
1	N	N	N	N	1
2	Y	N	N	N	2
3	N	Y	N	N	3
4	Y	Y	N	N	4
5	N	N	Y	N	5
6	Y	N	Y	N	6
7	N	Y	Y	N	7
8	Y	Y	Y	N	8
9	N	N	N	Y	9
10	Y	N	N	Y	10
11	N	Y	N	Y	11
12	Y	Y	N	Y	12
13	N	N	Y	Y	13
14	Y	N	Y	Y	14
15	N	Y	Y	Y	15
16	Y	Y	Y	Y	16



Multiple Choice Questions (with multiple choices)
Write Your Question Here

Q/What is the benefits of Networks?

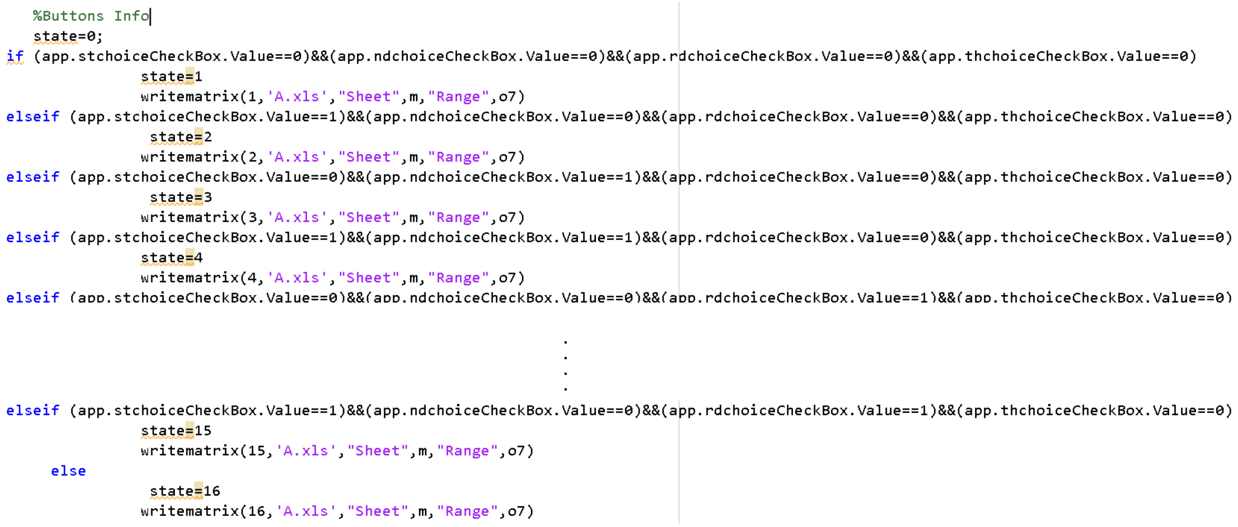
☑ 1st. choice Sharing info.

☑ 2nd choice Sharing resou

☑ 3rd choice Central Manaş

□ 4th choice Security

Figure 6: Multiple choice questions (with multiple choices)

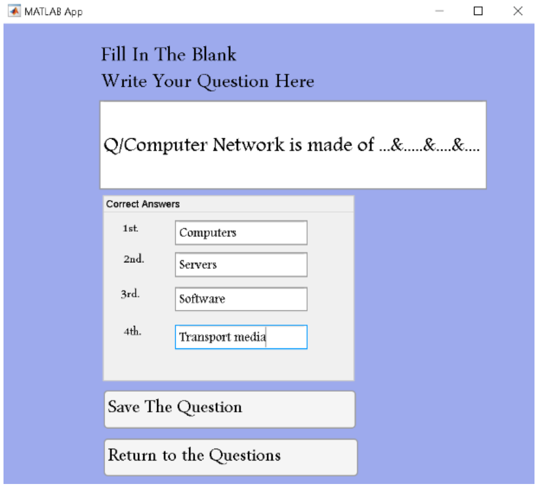


```
   %Buttons Info
   state=0;
if (app.stchoiceCheckBox.Value==0)&&(app.ndchoiceCheckBox.Value==0)&&(app.rdchoiceCheckBox.Value==0)&&(app.thchoiceCheckBox.Value==0)
              state=1
              writematrix(1,'A.xls',"Sheet",m,"Range",o7)
elseif (app.stchoiceCheckBox.Value==1)&&(app.ndchoiceCheckBox.Value==0)&&(app.rdchoiceCheckBox.Value==0)&&(app.thchoiceCheckBox.Value==0)
               state=2
              writematrix(2,'A.xls',"Sheet",m,"Range",o7)
elseif (app.stchoiceCheckBox.Value==0)&&(app.ndchoiceCheckBox.Value==1)&&(app.rdchoiceCheckBox.Value==0)&&(app.thchoiceCheckBox.Value==0)
               state=3
              writematrix(3,'A.xls',"Sheet",m,"Range",o7)
elseif (app.stchoiceCheckBox.Value==1)&&(app.ndchoiceCheckBox.Value==1)&&(app.rdchoiceCheckBox.Value==0)&&(app.thchoiceCheckBox.Value==0)
              state=4
              writematrix(4,'A.xls',"Sheet",m,"Range",o7)
elseif (app.stchoiceCheckBox.Value==0)&&(app.ndchoiceCheckBox.Value==0)&&(app.rdchoiceCheckBox.Value==1)&&(app.thchoiceCheckBox.Value==0)

                                                          .
                                                          .
                                                          .
                                                          .

elseif (app.stchoiceCheckBox.Value==1)&&(app.ndchoiceCheckBox.Value==0)&&(app.rdchoiceCheckBox.Value==1)&&(app.thchoiceCheckBox.Value==0)
              state=15
              writematrix(15,'A.xls',"Sheet",m,"Range",o7)
      else
               state=16
              writematrix(16,'A.xls',"Sheet",m,"Range",o7)
```

Code 5: MCQ (multiple choices)



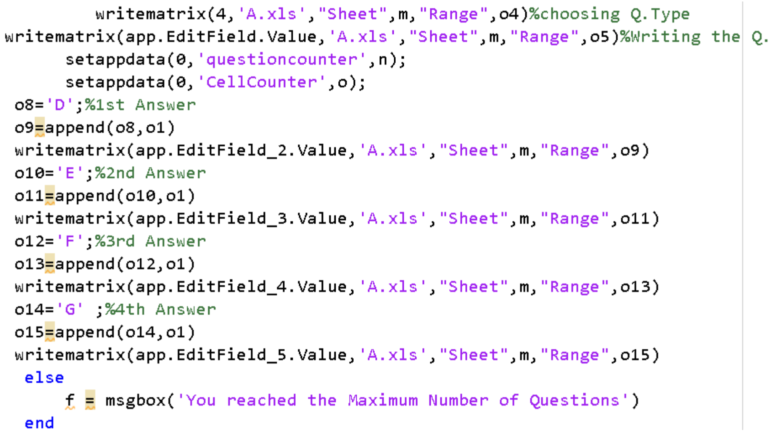
Fill In The Blank
Write Your Question Here

Q/Computer Network is made of ...&.....&....&....

Correct Answers

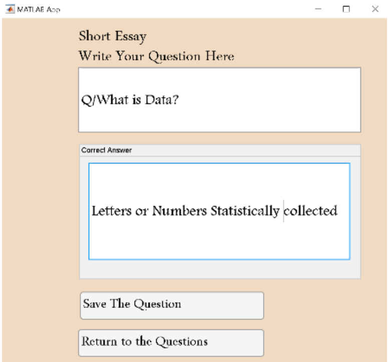
1st. Computers
2nd. Servers
3rd. Software
4th. Transport media

Figure 7: Fill in the blank Questions.



```
         writematrix(4,'A.xls',"Sheet",m,"Range",o4)%choosing Q.Type
writematrix(app.EditField.Value,'A.xls',"Sheet",m,"Range",o5)%Writing the Q.
      setappdata(0,'questioncounter',n);
      setappdata(0,'CellCounter',o);
 o8='D';%1st Answer
 o9=append(o8,o1)
 writematrix(app.EditField_2.Value,'A.xls',"Sheet",m,"Range",o9)
 o10='E';%2nd Answer
 o11=append(o10,o1)
 writematrix(app.EditField_3.Value,'A.xls',"Sheet",m,"Range",o11)
 o12='F';%3rd Answer
 o13=append(o12,o1)
 writematrix(app.EditField_4.Value,'A.xls',"Sheet",m,"Range",o13)
 o14='G' ;%4th Answer
 o15=append(o14,o1)
 writematrix(app.EditField_5.Value,'A.xls',"Sheet",m,"Range",o15)
  else
      f = msgbox('You reached the Maximum Number of Questions')
  end
```

Code 6: Fill in the blank



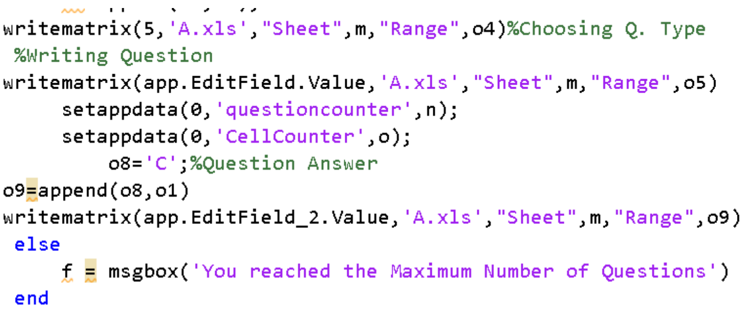
Short Essay
Write Your Question Here

Q/What is Data?

Correct Answer

Letters or Numbers Statistically collected

Figure 8: Essay question



```
writematrix(5,'A.xls',"Sheet",m,"Range",o4)%Choosing Q. Type
 %Writing Question
writematrix(app.EditField.Value,'A.xls',"Sheet",m,"Range",o5)
     setappdata(0,'questioncounter',n);
     setappdata(0,'CellCounter',o);
         o8='C';%Question Answer
o9=append(o8,o1)
writematrix(app.EditField_2.Value,'A.xls',"Sheet",m,"Range",o9)
 else
     f = msgbox('You reached the Maximum Number of Questions')
 end
```

Code 7: Essay questions.

spreadsheet is made as followed: The first cell (x,1) is the code of the question which is given the number 6, to the right a cell(x,2) to cell(x,5) is the sentence choices on the right, afterward the cells(x,6) to (x,9) contains their corresponding matching options on the left. This type of question is graded automatically. Match GUI is shown in Figure 9. The coding specifies the question type and gets the different options in the two columns presented in Code 8.

The Excel spreadsheet that will be generated corresponds to an exam named '**computer science**' with **6** questions (one for each type of question) with questions stated in Figures 4-9 presented in Figure 10.

2.3 Part Two: Opening an Existing Exam

The program can make a test for the user either for making a real CBT or for practicing with exams that are in the program's bank of questions. After clicking the opening an existing test is pressed, and a message box appears asking for the name of the user followed by another message box asking the user to specify that the test is a real CBT or a practice test, as shown in Figure 11. Coding such a task is presented in code 9.

The second message box input will have an impact on showing the button of 'show the correct answer', which will be invisible to the examinee taking a real CBT. Afterward, the user is asked to select a test from the existing available exams. The spreadsheet saved in the first part is turned into components in a list using the (listdlg) function. As illustrated in Figure 12 the coding for such a list is presented in Code 10.

Figure 13 shows the difference between a real CBT (a), and a practice test for the same exam (b), pressing the correct button will highlight the correct button in the autocorrect questions. (c) and (d). Coding (a) and (b) are presented in Code 11 (a), while the code for c is indicated in Code 11 (b).

The manually corrected questions (fill in the blank and essay) presented in Figure 14, and the "show the correct answer button" pressed and the questions with the correct answers are retrieved from the question spreadsheet and placed in the correct position.

The matching question will display the left and right sides of options randomly each time this type of question is invoked by an exam; with the correct match saved in the original exam. A click on the right side will cause a change in the color, and when the matching option on the left is clicked the same color appears and so on. Figure 15 shows the match question in a real exam.

Like creating a new exam, answering an exam (as a real exam or as practice) will create a new spreadsheet every time the program is used. The coding for the various question is given in Table 2 and Table3.

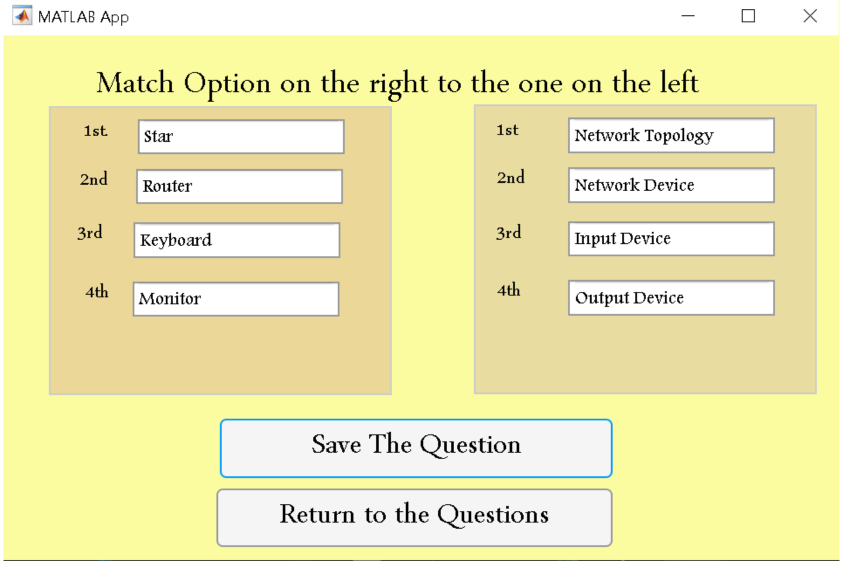
The answers are coded as shown in Table 3.

A spreadsheet is generated containing the student answers in Figure 16. This sheet is given the name of the examinee and contains the name of the exam taken as well as the different answers from the examinee's response to the different questions.

2.4 Part Three: Grading an Exam

This part is responsible for giving the proper grades to the examinee, the grading is semiautomatic. The fill in the blank and the essay questions are graded manually while the program automatically grades the other types. The pressing of the grade an exam button will display a list of the available answers, as shown in Figure 17.

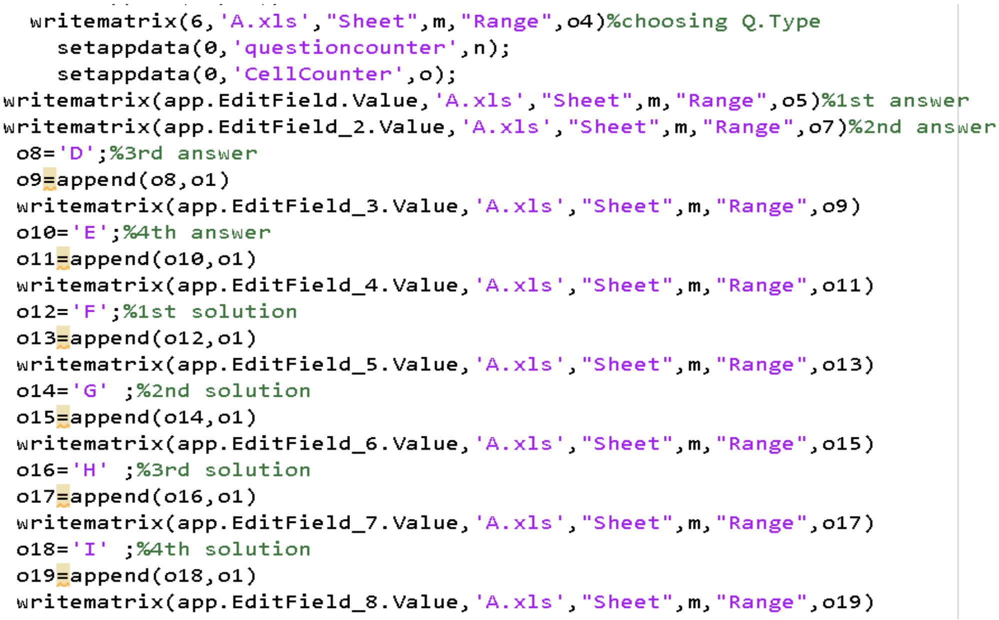
After choosing the desired exam, the questions appear in the order that they were saved with the examinee's answer plus the



Match Option on the right to the one on the left

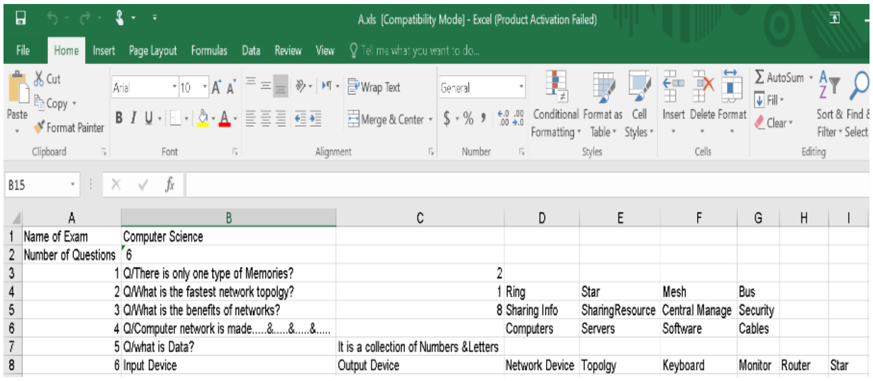
1st	Star	1st	Network Topology
2nd	Router	2nd	Network Device
3rd	Keyboard	3rd	Input Device
4th	Monitor	4th	Output Device

Figure 9: Match questions



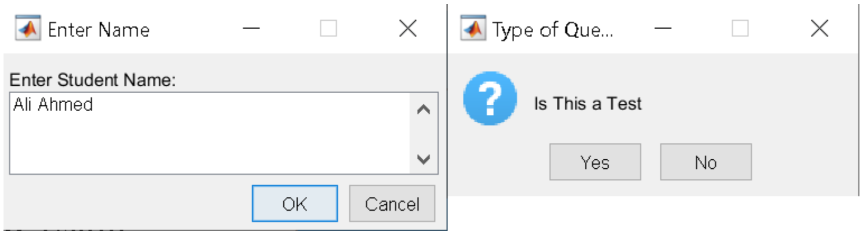
```
  writematrix(6,'A.xls',"Sheet",m,"Range",o4)%choosing Q.Type
    setappdata(0,'questioncounter',n);
    setappdata(0,'CellCounter',o);
writematrix(app.EditField.Value,'A.xls',"Sheet",m,"Range",o5)%1st answer
writematrix(app.EditField_2.Value,'A.xls',"Sheet",m,"Range",o7)%2nd answer
 o8='D';%3rd answer
 o9=append(o8,o1)
 writematrix(app.EditField_3.Value,'A.xls',"Sheet",m,"Range",o9)
 o10='E';%4th answer
 o11=append(o10,o1)
 writematrix(app.EditField_4.Value,'A.xls',"Sheet",m,"Range",o11)
 o12='F';%1st solution
 o13=append(o12,o1)
 writematrix(app.EditField_5.Value,'A.xls',"Sheet",m,"Range",o13)
 o14='G' ;%2nd solution
 o15=append(o14,o1)
 writematrix(app.EditField_6.Value,'A.xls',"Sheet",m,"Range",o15)
 o16='H' ;%3rd solution
 o17=append(o16,o1)
 writematrix(app.EditField_7.Value,'A.xls',"Sheet",m,"Range",o17)
 o18='I' ;%4th solution
 o19=append(o18,o1)
 writematrix(app.EditField_8.Value,'A.xls',"Sheet",m,"Range",o19)
```

Code 8: Match question



	A	B	C	D	E	F	G	H	I
1	Name of Exam	Computer Science							
2	Number of Questions	6							
3	1	Q/There is only one type of Memories?	2						
4	2	Q/What is the fastest network topolgy?	1	Ring	Star	Mesh	Bus		
5	3	Q/What is the benefits of networks?	8	Sharing Info	SharingResource	Central Manage	Security		
6	4	Q/Computer network is made.....&.....&.....&.....		Computers	Servers	Software	Cables		
7	5	Q/what is Data?	It is a collection of Numbers &Letters						
8	6	Input Device	Output Device	Network Device	Topolgy	Keyboard	Monitor	Router	Star

Figure 10: The generated spreadsheet after creating an exam



Enter Student Name:

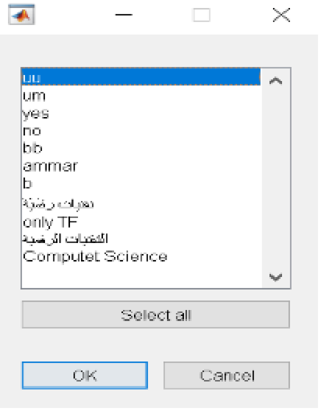
Ali Ahmed

Is This a Test

Figure 11: Name of the examiner and type of test message boxes

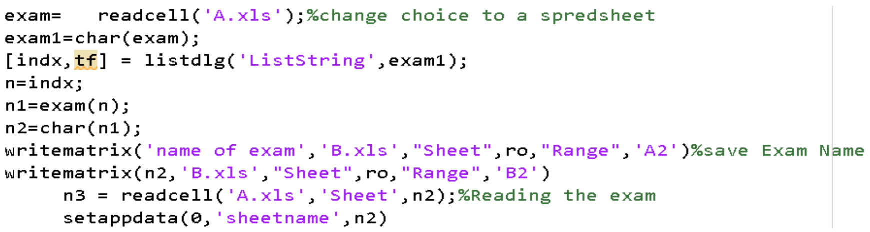
```
 %writing the new student name
prompt = {'Enter Student Name:'};
dlgtitle='Enter Name';
dims = [3 50];
definput = {''};
bo = inputdlg(prompt,dlgtitle,dims,definput);
 nameofstudent= "Name of Student";
 ro = char(bo);
 writematrix(nameofstudent,'B.xls',"Sheet",ro,"Range",'A1')
 writematrix(ro,'B.xls',"Sheet",ro,"Range",'B1')
 setappdata(0,'StudentName',ro);
%writing the name of the new student
 writematrix(ro,'B.xls','WriteMode',"append")
 %Determine if this is an exam or Practise
 flagexamOtrain =  questdlg('Is This a Test', ...
   'Type of Questions ', ...
   'Yes','No','');
  switch  flagexamOtrain
      case'Yes'
          bb=1
      setappdata(0,'ExamOTrain',bb);
      otherwise
          bb=2
      setappdata(0,'ExamOTrain',bb);
  end
```

Code 9: Dialog boxes



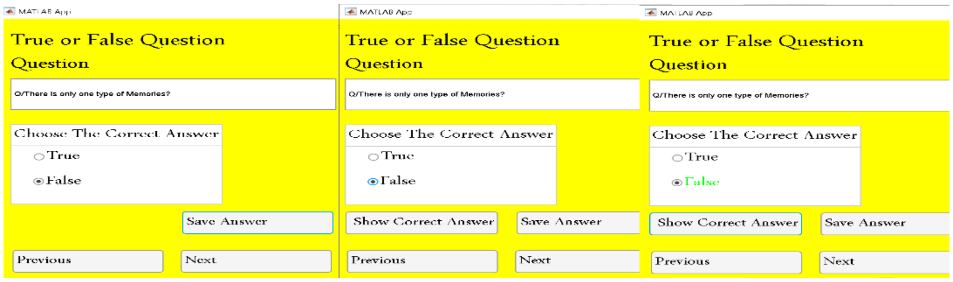
uu
um
yes
no
bb
ammar
b
مهارات رقمية
only TF
الكفايات الرقمية
Computet Science

Figure 12: Available exams list



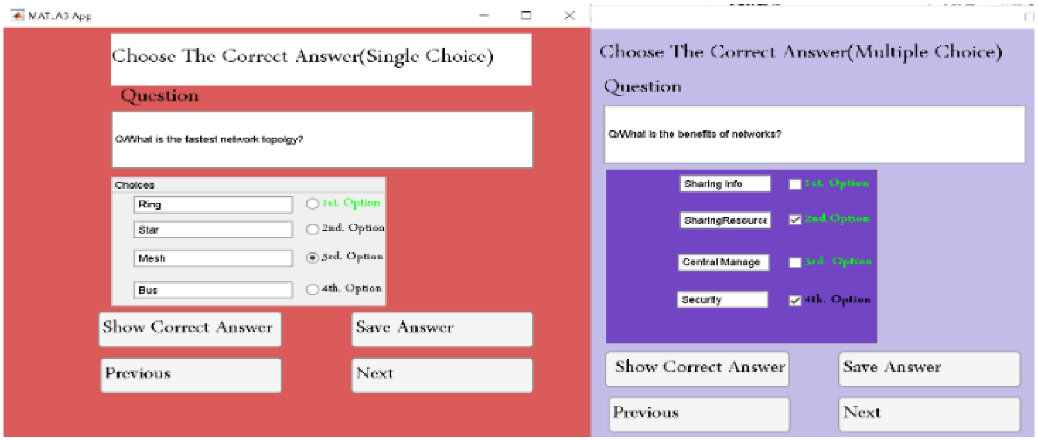
```
exam=   readcell('A.xls');%change choice to a spredsheet
exam1=char(exam);
[indx,tf] = listdlg('ListString',exam1);
n=indx;
n1=exam(n);
n2=char(n1);
writematrix('name of exam','B.xls',"Sheet",ro,"Range",'A2')%save Exam Name
writematrix(n2,'B.xls',"Sheet",ro,"Range",'B2')
     n3 = readcell('A.xls','Sheet',n2);%Reading the exam
     setappdata(0,'sheetname',n2)
```

Code 10: The dialog box for choosing an exam



Q/There is only one type of Memories?

(a) (b) (c)



Q/What is the fastest network topolgy?

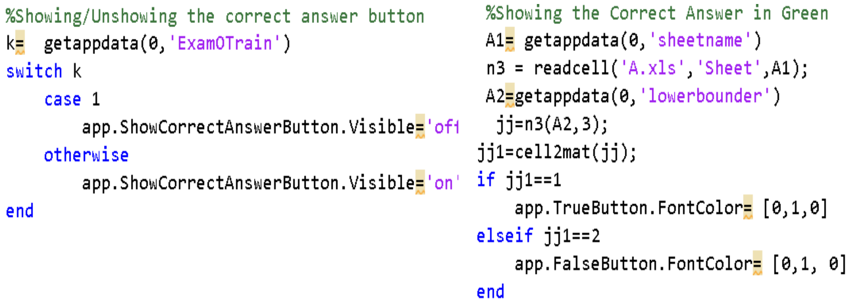
Ring ○1st. Option
Star ○2nd. Option
Mesh ●3rd. Option
Bus ○4th. Option

Q/What is the benefits of networks?

Sharing Info □1st. Option
SharingResource ☑2nd. Option
Central Manage □3rd. Option
Security ☑4th. Option

MATLAB App
Choose The Correct Answer(Single Choice)
Question
Choices
Choose The Correct Answer(Multiple Choice)
Question
Show Correct Answer
Save Answer
Previous
Next

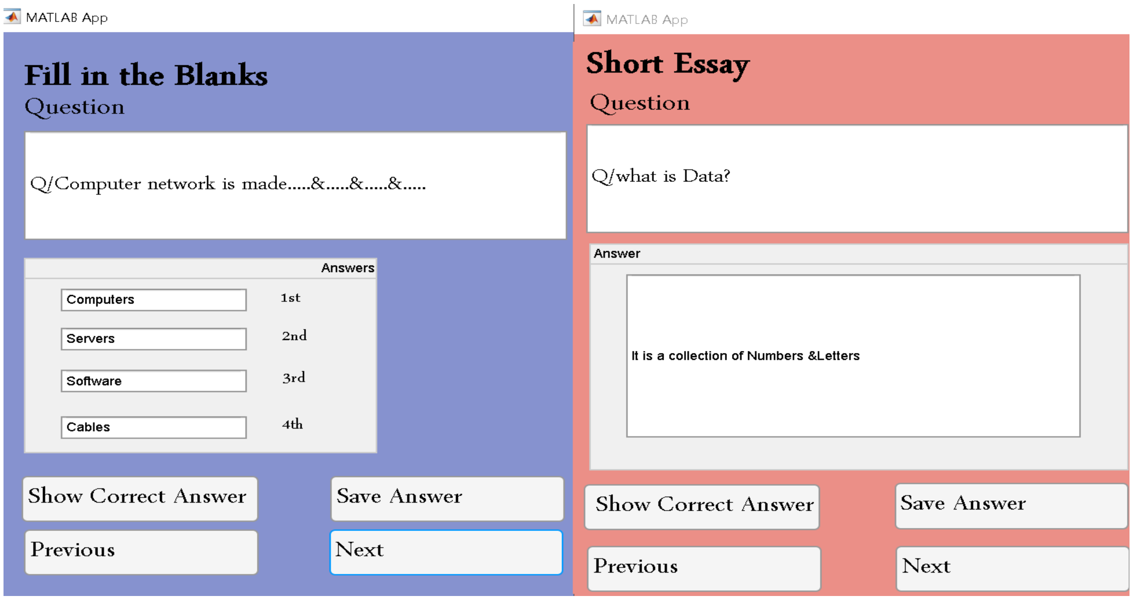
Figure 13: Real & practice tests (a) (b)



```
%Showing/Unshowing the correct answer button
k=  getappdata(0,'ExamOTrain')
switch k
    case 1
        app.ShowCorrectAnswerButton.Visible='off
    otherwise
        app.ShowCorrectAnswerButton.Visible='on'
end
```

```
%Showing the Correct Answer in Green
A1= getappdata(0,'sheetname')
n3 = readcell('A.xls','Sheet',A1);
A2=getappdata(0,'lowerbounder')
 jj=n3(A2,3);
jj1=cell2mat(jj);
if jj1==1
    app.TrueButton.FontColor= [0,1,0]
elseif jj1==2
    app.FalseButton.FontColor= [0,1, 0]
end
```

Code 11: T or F question with the ability to hide show the right answer



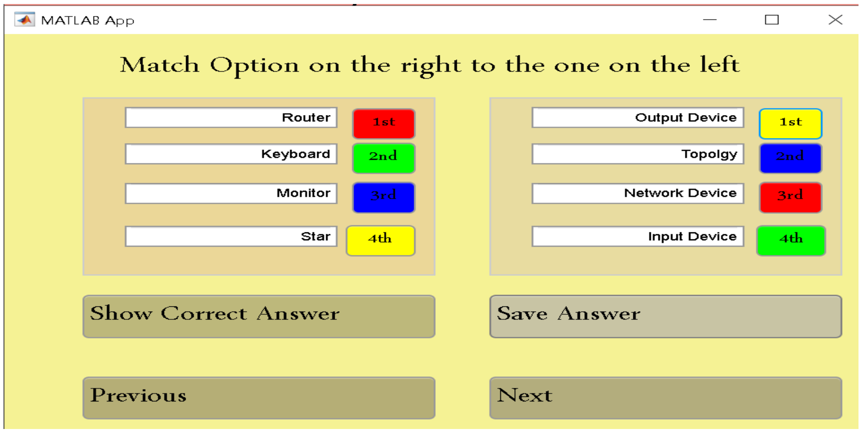
Q/Computer network is made.....&.....&.....&.....

Computers	1st
Servers	2nd
Software	3rd
Cables	4th

Q/what is Data?

It is a collection of Numbers &Letters

Figure 14: Showing the correct answer in manually corrected questions



Router	1st	Output Device	1st
Keyboard	2nd	Topolgy	2nd
Monitor	3rd	Network Device	3rd
Star	4th	Input Device	4th

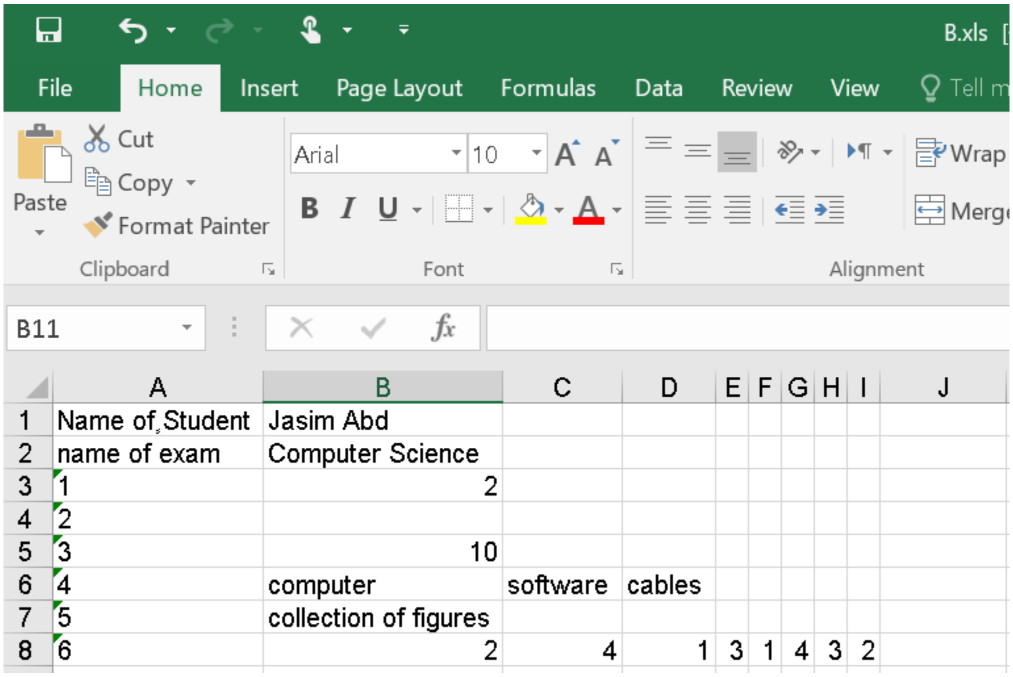
Figure 15: The Matching question in an exam

Table 2: The header for the answer

Cell name	contents		
A1	'Name of student'	B1	Actual name of the student
A2	'Name of exam'	B2	Actual name of the exam
A3…….An	Type of Question by number(1,2,3,4,5 or 6)	B3	The student's answer (might be True or False according to Table 3

Table 3: Coding the answers

Cell name	contents	Cell name	contents	
Ax	‘1:(for T or F question)	Bx	1 or 2	1: True 2: False
Ax+1	2 :(MCQ single choice)	Bx+1	1,2,3 or 4	1:1st. answer selected 2:2nd. Answer 3:3rd answer 4:4th.answer
Ax+2	3 :(MCQ multiple choice)	Bx+2	From 1 to 16	According to the state illustrated in Table (1)
Ax+3	4 :(Fill in the blank question)	Bx+3,Cx+3,Dx+3, Ex+3	Answers from 1 to 4 according to the question	
Ax+4	5:(Essay question)	Bx+4	Answer to the question	
Ax+5	6 :(Match question)	Bx+5,Cx+5,Dx+5, Ex+5,	The select items on the right are coded in numbers from 1 to 4	
		Fx+5,Gx+5, Hx+5,Ix+5	The select items on the left are coded in numbers from 1 to 4	



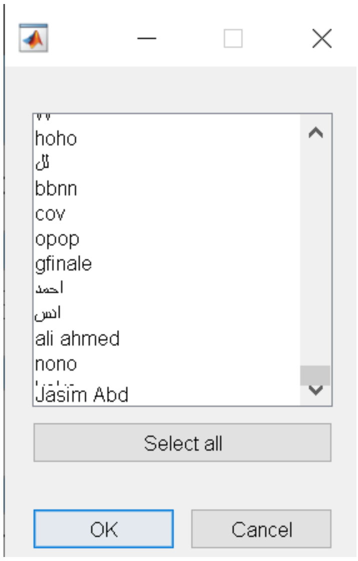
	A	B	C	D	E	F	G	H	I	J
1	Name of Student	Jasim Abd								
2	name of exam	Computer Science								
3	1	2								
4	2									
5	3	10								
6	4	computer	software	cables						
7	5	collection of figures								
8	6	2	4	1	3	1	4	3	2	

Figure 16: The spreadsheet was generated from the examinee’s answer

correct answer (for the autocorrect questions). The left side contains the name of the student, exam name, total score points, and score for an individual question (a spinner component that has a default number set to the grade number for a question. This number can be changed by the exam corrector if he wants to change the mark for the auto-graded question or give the proper degree for the fill in the blank or essay.

After giving the proper degree “Degree accreditation button” is pressed, and the total degree is calculated, as shown in Figure 18.

For the manual grading questions, the spinner component is used to select the proper mark for the question. When the final question is reached and this button is pressed followed by the “Degree accreditation button” the final mark and the name of



hoho
كل
bbnn
cov
opop
gfinale
احمد
انس
ali ahmed
nono
Jasim Abd

Figure 17: List of the available examinee

the student is stored in a third spreadsheet.

3 Conclusion

Information and communication technology (ICT) advancements will improve community life and allow governments to operate more efficiently and sustainably. The CBT is one of the cornerstones of any LMS. From the proposed system, we can draw the following conclusions:

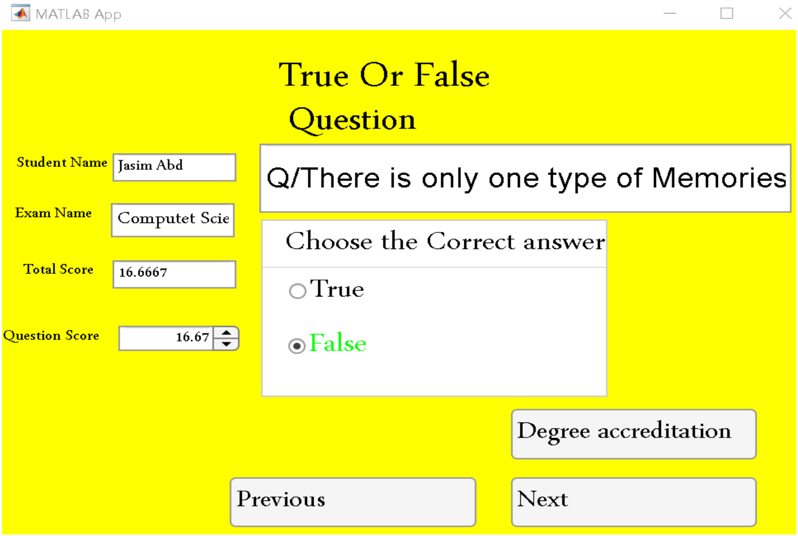
1-The three essential steps for the system (creating, taking, and grading) are made by MATLAB in conjunction with Microsoft Excel so that each question can be traced thoroughly and clearly in all three steps no matter the size of the exam. The debugging was made for each type of question separately with no margin of error.

2-There are six types of questions in the program chosen according to the frequency of appearance in the major LMS and CBT programs. Many other common types of questions like re-order and Hot spot can be easily added to the program.

3- Many new types of questions can be invented and added to the program, exploiting MATLAB capabilities. For example, MATLAB Simulink is one of the most promising tools to be used for a real-time simulation in all fields of technology and creating questions about these situations.

4-Although the exams are predefined(static) because the program is mainly designed to be used in real exams; each participating student has equal opportunity questions, further analysis of the answers can be made, to measure the response patterns and use that to tweak the questions. This topic is worth further investigation and can be used in future work.

5-The program interfaces are designed in the English language and they can be easily changed to other left-to-right languages simply by changing the interfaces for each question. On the other hand, right-to-left needs an added effort to change the alignment of the text boxes to use these languages in the program.



True Or False Question

Student Name: Jasim Abd

Exam Name: Computet Scie

Total Score: 16.6667

Question Score: 16.67

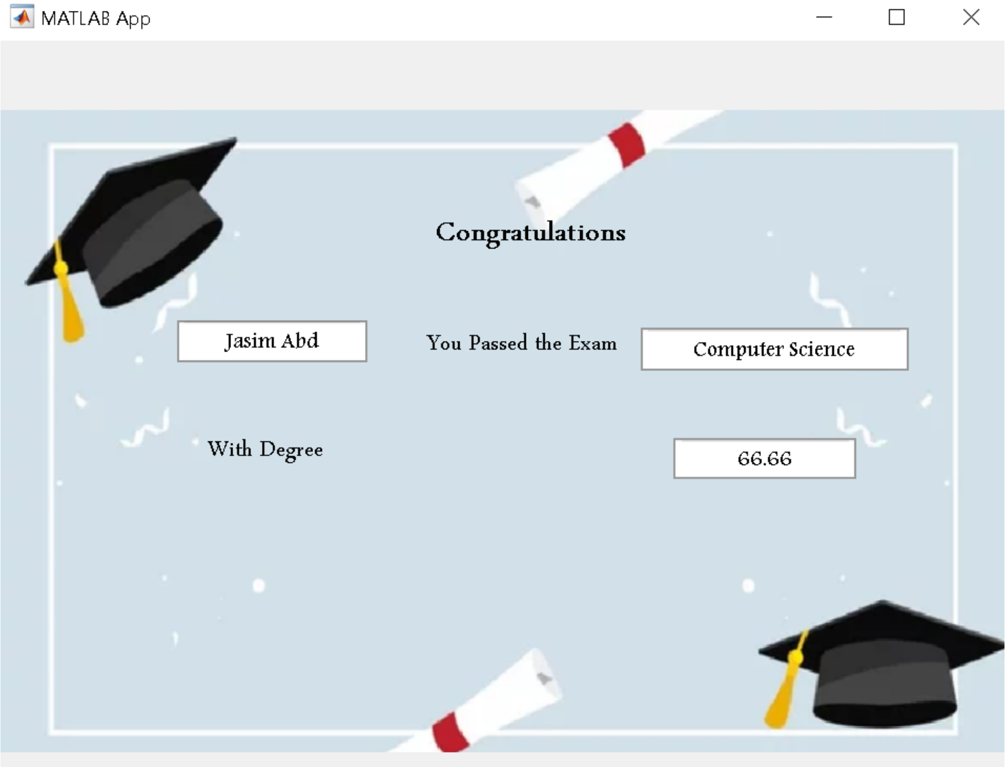
Q/There is only one type of Memories

Choose the Correct answer

○ True

● False

Figure 18: T or F question grading



Congratulations

Jasim Abd

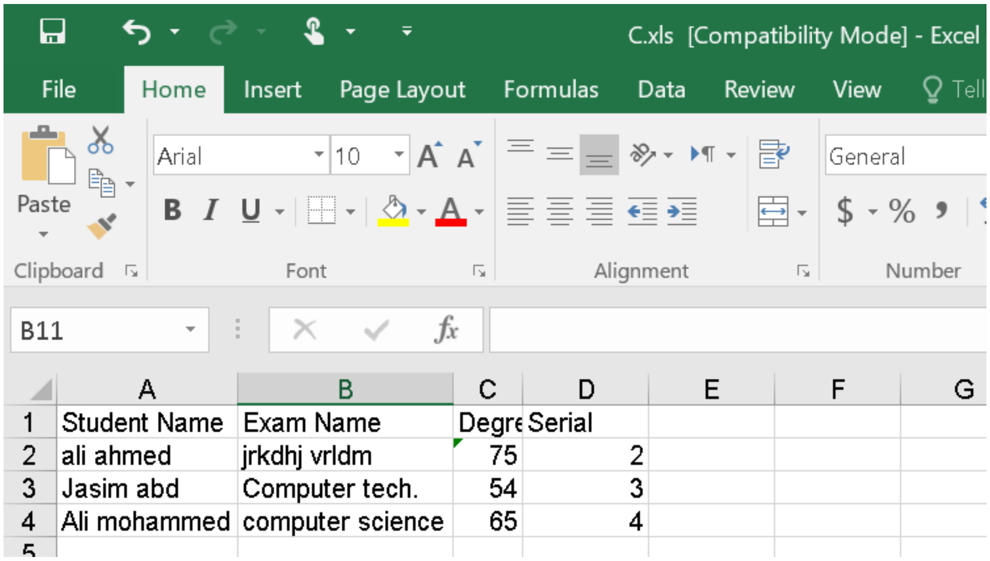
You Passed the Exam

Computer Science

With Degree

66.66

Figure 19: Pass exam certification



	A	B	C	D	E	F	G
1	Student Name	Exam Name	Degre	Serial			
2	ali ahmed	jrkdhj vrldm	75	2			
3	Jasim abd	Computer tech.	54	3			
4	Ali mohammed	computer science	65	4			
5							

Figure 20: Record of examinee

References

[1] R. Abdulhussien and H. Najeeb, "Improving Measurement of Effectiveness of Blended Learning in Iraqi Education Using SVM," *Iraqi Journal of Science*, 63(9):4057-4066, 2022.

[2] M. Ajinaja, "The Design and Implementation of a Computer Based Testing System Using Component-Based Software Engineering," International Journal of Computer Science and Technology, 8(9):58-65, March 2017.

[3] A. Baghdadi, "The Computer Based Tests: A Digital

Substitution for the Iraqi Postgraduate Students," *Ibn Al-Haitham Jour. for Pure & Appl. Sci*., 31(3):171-177, 2018, https://doi.org/10.30526/31.3.2019.

[4] A. Baghdadi, "An Automatic System to Grade Multiple Choice Questions Paper-Based Test." *J. of Al-Anbar University for Pure Science*, 3(1):174-181, 2009.

[5] O. Chibuzo and D. Isiaka, "Design and Implementation of Secure Browser for Computer-Based Tests," *International Journal of Innovative Science and Research Technology*, 5(8):1347-1356, August 2020.

[6] N. Dogan, N. UYSAL, and R. Hambelton, "An Overview of E-Assessment Hacettepe University Journal of Education," *University Journal of Education*, 35(Special Issue):1-5, 2020, doi: 10.16986/HUJE.2020063669.

[7] S. Ejim, "An Overview of Computer Based Test", DOI: 10.13140/RG.2.2.32040.88326, February 2017.

[8] A. Evwiekpaefe and V. Eyinla, "Implementing Fingerprint Authentication in Computer-Based Tests," *Nigerian Journal of Technology*, 40(2):284-291, 2021.

[9] M. Jodin, "Measurement Efficiency of Innovative Item Formats in Computer-Based Testing," *Journal of Educational Measurement*, 40(I):1-1s, 2003.

[10] O. Kravchenko, N. Shelenkova, M. Shelenkova, and I. Boichevska, "Computer-Based Testing Complex "Profvybir": Occupational Guidance Diagnostics Journal of Physics: Conference Series, Volume 1828, 2020 International Symposium on Automation, Information and Computing (ISAIC 2020) 2-4 December 2020, Beijing, China. doi:10.1088/1742-6596/1828/1/012125.

[11] R. Luecht and S. Sireci S "A Review of Models for Computer-Based Testing," College Board Research Report, 2011-12.

[12] K. Maneekhao N. Jaturapitakkul R. Watson R., and S. Tepsuriwong., "Developing an Innovative Computer-Based Test," *Prospect*, 21(2):34-46, August 2006.

[13] Naseef Husam Mohammad, Nada Thanoon Ahmed, and Yasmin Makki Mohialden, "Development of Multiple Computer-Based Testing System Using Open-Source Programing Model," Journal of Physics: Conference Series 1804 012063, 2021, 012063 IOP Publishing doi:10.1088/1742-6596/1804/1/012063.

[14] H. Nurhikmah, H. Abdul Gani, M. Pratama, and H. Wijaya, "Development of an Android-Based Computer Based Test (CBT) In Middle School," *Journal of Education Technology*, 5(2):272-281, 2021.

[15] E. Nweneka, "A Secure Online Computer-Based Test System Using Facial Recognition Biometric Authentication," A Case Study of Mountain Top University, 2021.

[16] F. Okocha, "Student Perception of Computer-Based Testing in Kwara State, Nigeria," *International Journal of Web-Based Learning and Teaching Technologies,* 17(1):1-11, 2022. DOI: 10.4018/IJWLTT.294575.

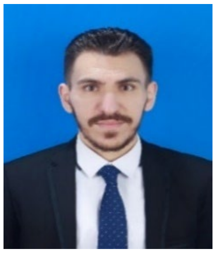
[17] P Patel and A. Karkare, "Accessibility Evaluation of Computer Based Tests," arXiv:1905.01825v1, 2019.

[18] M. Umar and A. Jaya, "The Bimasoft Application as Computer Based Test (CBT) Learning Evaluation Media: An Analysis of the Effectiveness Using Android," *Jurnal Pengkajian Ilmu dan Pembelajaran Matematika dan IPA IKIP Mataram*, 10**:**821-830, 3 July 2022.

[19] M. Schurig, J. Jungjohann, and M. Gebhardt2, "Minimization of a Short Computer-Based Test," Reading Frontiers in Education, 6:1-12 | Article 684595 2021.

Baghdadi Ammar Awni Abbas is an Assistant Professor of information technology. He received his Ph.D. from Kharkiv in Radioelectronics/Ukraine in 2015. He has over 30 published papers in 3 languages.

He is now a member in the Faculty of the Department of Radio and Television/College of mass media/University of Baghdad. His lifetime devotion is programming and teaching others how to program. MATLAB and C language are his favorite tools where he used these tools in image processing, computer vision pattern recognition, human-computer interaction, and artificial intelligence. He is currently working on a large project of interactive lectures for undergraduate students.



Mohammed Al-Mukhtar received his M.Sc. degree in Information and Communication Engineering from Huazhong University of Science and Technology (HUST), Wuhan, China, in 2019. Currently, he is working as a Lecturer at the University of Baghdad. His research interests include Computer Application, Medical Image Analysis with deep learning techniques, Computer Vision, Object Recognition, and Object Segmentation.

Journal Submission

The International Journal of Computers and Their Applications is published four times a year with the purpose of providing a forum for state-of-the-art developments and research in the theory and design of computers, as well as current innovative activities in the applications of computers. In contrast to other journals, this journal focuses on emerging computer technologies with emphasis on the applicability to real world problems. Current areas of particular interest include, but are not limited to: architecture, networks, intelligent systems, parallel and distributed computing, software and information engineering, and computer applications (e.g., engineering, medicine, business, education, etc.). All papers are subject to peer review before selection.

A. Procedure for Submission of a Technical Paper for Consideration

1. Email your manuscript to the Editor-in-Chief, Dr. Ajay Bandi. Email: ajay@nwmissouri.edu.
2. Illustrations should be high quality (originals unnecessary).
3. Enclose a separate page (or include in the email message) the preferred author and address for correspondence. Also, please include email, telephone, and fax information should further contact be needed.
4. **Note**: Papers shorter than 10 pages long will be returned.

B. Manuscript Style:

1. **WORD DOCUMENT**: The text should be **double-spaced** (12 point or larger), **single column** and **single-sided** on 8.5 X 11 inch pages. Or it can be single spaced double column.

   LaTex DOCUMENT: The text is to be a double column (10 point font) in pdf format.
2. An informative abstract of 100-250 words should be provided.
3. At least 5 keywords following the abstract describing the paper topics.
4. References (alphabetized by first author) should appear at the end of the paper, as follows: author(s), first initials followed by last name, title in quotation marks, periodical, volume, inclusive page numbers, month and year.
5. The figures are to be integrated in the text after referenced in the text.

C. Submission of Accepted Manuscripts

1. The final complete paper (with abstract, figures, tables, and keywords) satisfying Section B above in **MS Word format** should be submitted to the Editor-in-Chief. If one wished to use LaTex, please see the corresponding LaTex template.
2. The submission may be on a CD/DVD or as an email attachment(s). **The following electronic files should be included:**
   - Paper text (required).
   - Bios (required for each author).
   - Author Photos are to be integrated into the text.
   - Figures, Tables, and Illustrations. These should be integrated into the paper text file.
3. Reminder: The authors photos and short bios should be integrated into the text at the end of the paper. All figures, tables, and illustrations should be integrated into the text after being mentioned in the text.
4. The final paper should be submitted in (a) pdf AND (b) either Word or LaTex. For those authors using LaTex, please follow the guidelines and template.
5. Authors are asked to sign an ISCA copyright form (http://www.isca-hq.org/j-copyright.htm), indicating that they are transferring the copyright to ISCA or declaring the work to be government-sponsored work in the public domain. Also, letters of permission for inclusion of non-original materials are required.

Publication Charges

After a manuscript has been accepted for publication, the contact author will be invoiced a publication charge of **$500.00 USD** to cover part of the cost of publication. For ISCA members, publication charges are **$400.00 USD** publication charges are required.

Revised 2020

ISCA INTERNATIONAL JOURNAL OF COMPUTERS AND THEIR APPLICATIONS Vol. 30 No. 3, Sept. 2023