IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 October 2024, accepted 26 November 2024, date of publication 9 December 2024, date of current version 10 January 2025.

Digital Object Identifier 10.1109/ACCESS.2024.3514079

RESEARCH ARTICLE

MathVision: An Accessible Intelligent Agent for Visually Impaired People to Understand Mathematical Equations

MUHAMMAD AWAIS AHMAD[1], TAUQIR AHMED[1], MUHAMMAD ASLAM[1], AMJAD REHMAN[2], (Senior Member, IEEE), FATEN S. ALAMRI[3], SAEED ALI BAHAJ[4], AND TANZILA SABA[2], (Senior Member, IEEE)

[1]Department of CS, University of Engineering and Technology, Lahore, Punjab 39161, Pakistan
[2]Artificial Intelligence and Data Analytics Lab, CCIS, Prince Sultan University Riyadh 11586, Saudi Arabia
[3]Department of Mathematical Sciences, College of Science, Princess Nourah Bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia
[4]MIS Department College of Business Administration, Prince Sattam Bin Abdulaziz University, AlKharj 11942, Saudi Arabia

Corresponding author: Faten S. Alamri (fsalamri@pnu.edu.sa)

This research was funded by Princess Nourah bint Abdulrahman University and Researchers Supporting Project number (PNURSP2025R346), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

ABSTRACT 2.2 billion people worldwide suffer from some form of vision impairment, according to the World Health Organization. Children with vision impairment and visual impairment may experience impaired physical, linguistic, and cognitive development, resulting in reduced levels of academic accomplishment. Many visually impaired people are working in the education sector whether they are students or teachers. Without external assistance reading of mathematical equations in images for visually impaired people is very challenging due to the complexity of notations, symbols, and variables. This paper presents a model named MathVision which converts the mathematical equation into voice. This voice is quite helpful for visually impaired people to understand mathematical equations. The proposed model is comprised of YOLOv7 object detection architecture to detect and categorize mathematical equations inside images into four distinct types: limits, trigonometry, integration, and an additional category. The input image is divided into a grid by the YOLOv7 model, and each grid cell is responsible for finding equations that fall into its respective category. bounding box coordinates, object labels, and probability scores are predicted for each equation. In the next stage, a fine-tuned DenseNet is utilized for detailed feature extraction from mathematical equation images. This involves optimizing a pre-trained DenseNet model to capture intricate patterns specific to equations. The fine-tuned DenseNet enhances overall accuracy in equation detection and categorization within the system. In the subsequent phase, an attention mechanism-based LSTM network is employed to generate natural language descriptions for mathematical equations. During the decoding process, the model is better able to focus on pertinent portions of the equation due to the integration of attention. The LSTM architecture, chosen for its effectiveness with sequential data, is trained on a dataset containing paired examples of equations and corresponding human-generated descriptions. Fine-tuning includes optimizing hyperparameters for the task, and evaluation metrics such as the BLEU score are used to assess the model's performance in generating accurate and contextually relevant textual representations for the detected mathematical content. Our text-to-speech system takes input in the form of a natural language sentence generated by the LSTM model and converts it to the voice. This TTS using natural language processing analyzes and processes the text then it converts this processed text into speech using digital signal processing technology. A platform-independent pyttsx3 python library is used for converting text into speech. It also works offline which is the main reason for using this library in this research work. As there was

The associate editor coordinating the review of this manuscript and approving it for publication was Rajeeb Dey.

no dataset available of mathematical equations with their natural language description, we created a custom dataset. We conducted real-world experiments in various visually impaired schools to see whether visually impaired students can understand mathematical equations by hearing the voice. These experiments prove that the MathVision Model is an efficient way for visually impaired students to read and write mathematical equations by listening to the voice of equations generated by proposed model.

INDEX TERMS Mathematical equations, fine-tuned, YOLO v7, convolution neural network, attention mechanism, long short term memory, neural text to speech, technological development.

I. INTRODUCTION

2.2 billion people worldwide suffer from some form of vision impairment, according to the World Health Organization. Children with vision impairment and visual impairment may experience impaired physical, linguistic, and cognitive development, resulting in reduced levels of academic accomplishment [1]. Every stage of a student's academic career requires them to learn mathematics. An alternate method for developing and enhancing their mathematics abilities is through solving mathematical questions [2]. As visually impaired students are limited to reading and writing they find it particularly challenging to learn mathematics [3]. Human readers usually assist those types of students in accessing and understanding the mathematical equations in various documents. The availability of human readers for that kind of student at every time is impossible due to cost and a limited number of trained individuals. For the visually impaired, Braille is the preferred and most practical method of accessing documents [4]. Unfortunately, braille covers limited documents, as braille does not cover the conversion of mathematical equations due to the complexity of various symbols and operators. Additionally, students who use literary transcription of braille it is challenging for them. Sound-based representation of documents is also a widely used way of accessing information for visually impaired students [5]. In this regard, various audio material is used for studying content talking books and daisy books are examples of that kind of audio material. However, these books are also unable to cover the mathematical equations. Students who are visually impaired or have visual impairments (VI) frequently use text-to-speech (TTS) methods to read electronic content on computers [6]. TTS systems generate synthetic speech from digital text. Unfortunately, most of the available TTS systems are only capable of reading plain text. When an image of a mathematical equation is given as an input to available TTS systems, they are unable to generate voice. Many academics understood how significant it is to improve the readiness of mathematical content for visually impaired students, therefore text to text-to-speech-based mathematical equation reading systems were introduced [7].

To read a mathematical equation from an image, some of these advanced systems need additional words. The rendered audio is lengthy due to the presence of extra words. As a result of the lengthy audio duration, students may not always be able to understand the internal meaning of mathematical equations. Additionally, most of this intelligent mathematics reading system accept mathematical equations in LaTex form or other comparable markup languages [8]. The documents that contain mathematical equations, unfortunately, are not in the LaTex format as mathematical equations can also be found in images [9]. Therefore, it's challenging to generate a LaTex sequence of corresponding mathematical equations. This research aims to develop a framework called MathVision which converts the mathematical equation present in an image to voice which gives a clear understanding of mathematical equation to visually impaired students. Reading mathematical equations somehow relates to image caption problems in computer vision [10]. The proposed MathVision model generates the voice of mathematical equations.

This voice helps visually impaired people to understand mathematical equations quite comfortably. For Example, $\int cosx\,dx$ given as input to proposed model first detects the equation and classifies that the given equation belongs to the integration class of mathematical equation, then feature extracted and these features are passed to the next module of proposed model that generates natural language description and at last, this natural language sentence is converted to voice. However, reading of mathematical equation from an image to conversion into voice is very challenging as mathematical equations contain various symbols, operators, and their changing positions [11].

As per our knowledge, there is no existing work that generates the voice of mathematical equations. The main motivation behind our work is to facilitate the visually impaired community working in the discipline of mathematics. So that they can understand mathematical equations as normal human being reads. In this paper, we propose a novel model named MathVision to generate voices for the mathematical equations which gives a depth understanding of mathematical equations. Proposed framework comprises four modules: detection and classification, encoder, decoder, and TTS. YOLOv7 (you only look once) is used for the detection and classification of mathematical equation images [12]. A convolution neural network [13] is used as an encoder to encode the mathematical equation images. We have trained long short-term memory which is responsible for receiving extracted features from the encoder and then generating the natural language description [14].

The attention mechanism is also used with the decoder so that it can focus on specific parts of an image. This natural language sentence acts as the input of the TTS module [15]

which converts it to voice. We named this model as MathVision model.

The main contributions of this paper are as follows:

- A dataset that contains four categories of mathematical equations is built with their natural language description.
- We presented a novel model MathVision which generates voices of mathematical equations.
- We conducted real-world experiments to validate the effectiveness of our MathVision Model for providing a better understanding of mathematical equations.
- We also measure the performance of our long short-term memory (LSTM) module which generates natural language descriptions of mathematical equations.

The rest of the document is organized as follow: Section II presents the related work whereas Section III describes the proposed model. Section IV provides information about the collected dataset and Section V discusses the results. The research contribution are given in Section VI, future work are written in Section VII and finally conclusions are presented in Section VIII.

II. RELATED-WORK

Zhao et al. [16] proposes a novel approach for handwritten mathematical equation recognition. The author used a bidirectional trained transformer. The authors note that traditional methods for recognizing handwritten mathematical expressions have limitations due to the complexity and variability of such expressions. The proposed approach uses a transformer-based model that can learn the context of handwritten expressions by incorporating bidirectional training. Tope et al. [17] have explored seven methods but the convolutional neural network has the highest accuracy among all of them and convolution neural network (CNN) proved to be the best algorithm for recognition of handwritten mathematical expressions. Ogwok and Ehlers [18] implemented a model for recognizing and contextualizing the mathematical expression from noisy images.

The system uses the processing of images and neural networks to achieve its objective. The proposed system tries to reduce the issues that come at the segmentation stage and recognition. The authors proposed an idea in [19] to recognize handwritten mathematical equations and evaluate them offline using the CNN algorithm for classification. In the first step, the scanned worksheet is sent to the work-spaces detection module. From the given worksheet, this particular module returns to a rectangular workspace. Then, these workspaces are passed through to a module named line extraction that extracts all the lines. In the next step, these extracted lines are then passed to the character segmentation module. In this step, segmentation of characters takes place. Using a deep learning model, the characters are classified. Finally, the evaluation module analyzes the line, and a green/red bounding box is drawn depending upon whether the line is correct or not. Agarwal et al. [20] summarized the important aspects of the past work. The research also investigated different spectrums of approaches in mathematical expression recognition (MER).

The domains of machine translation and image captioning face similar issues in terms of integrating diverse feature modalities, improving markup embedding, and eliminating visual parsing ambiguities. Sakshi and Kukreja [21] provided a detailed review which plays an important role in the interesting extractions about MER representation. A deeper look into aspects of mathematical recognition methodology is offered by this report statistically as well as graphically. Neoteric recognition models and approaches are also emphasized in this paper. It marks a significant contribution by identifying, extracting, and then categorizing contemporary recognition techniques. Based on the frequency of occurrence, various recognition approaches have been emphasized, and all associated ideas have been depicted graphically.

In reference [22] a framework that uses an attention-based encoder and decoder is used for the recognition of mathematical equations (Equations). Symbol prediction and structuring phasing of a mathematical equation are also included in this study. The purpose of the encoder is to take out features from the image received as input while the decoder is used for the prediction of one symbol at each time step that leads towards outputting a sequence. Detection and recognition of equations are integrated into a unified system by Pong et al. [23]. Document images are given as input to that system, and it outputs the latex strings representing the detected equations in the document. These are detected by the deep learning algorithm named YOLOV3 then these images are fed into watch attend and parsed for recognition

Zhang et. al [24] used a CNN and Recurrent Neural Network (RNN) based approach that is used for optimizing the processing of formula symbols in mathematical expressions. In this proposed method, image features are extracted using CNN and then RNN is used to generate mathematical expressions. Using IM2Latex 100K data set accuracy of 88.42% is achieved. The mathematical equation description (MED) framework is proposed by Mondal and Jawahar [14]. Visually impaired and visually impaired students can take help from this framework for reading and interpreting the hidden meaning of mathematical equations (Equations) in document images. The objective of the research is achieved using two modules: ME images are encoded using a convolution neural network. This module is known as the encoder. Another module is named as a decoder that uses long short-term memory. This module takes the intermediate representation of ME images and their natural language description is generated. This research work used two different data sets their corresponding textual description is generated.

An offline ME recognition system that can efficiently recognize equations, this system uses a technique known as the generative vision model [25]. This model has a quality that it can learn and generalize from a few examples. The 20 images that are taken from the MNIST images dataset

obtain 60 percent symbol-level identification accuracy. Three aspects of the proposed system are evaluated in the research (i) Accuracy of segmentation (ii) recognition accuracy at the symbol level (iii) recognition accuracy at expression level. Wu et al. [26] propose a new approach for online handwritten mathematical equation recognition that is accurate and interpretable the proposed approach uses a graph-to-graph neural network that can learn the structural relationships between the different symbols and sub-expressions that make up a mathematical expression. The model is trained on a large dataset of annotated mathematical expressions and is designed to be both accurate and interpretable, allowing users to understand how the model arrives at its predictions.

The proposed approach in [27] uses two separate models, one for recognizing the individual symbols in a handwritten expression and another for recognizing the overall structure of the expression. The two models are trained using a bi-directional mutual learning algorithm, where the output from one model is used to inform the training of the other model. The authors also introduce an attention aggregation mechanism for performance improvement of the recognition model. The survey conducted by paper [29] covers a wide range of topics, including preprocessing techniques, feature extraction, recognition models, and applications of the recognition models. The authors provide an overview of the different types of data used for training and testing recognition models, such as isolated symbols, isolated characters, and full expressions. The authors also discuss the different types of recognition models used for online handwritten mathematical expression recognition, including support vector machines, deep learning algorithms, and Hidden Markov algorithms. Shinde Sagar et al. [28] propose an approach based on machine learning for handwritten mathematical equations and symbol recognition.

The authors experiment using various machine learning algorithms like K-Nearest Neighbour (KNN), Support Vector Machine (SVM), and Convolutional Neural Networks (CNNs). Various evaluation metrics are used by the author to measure the working of each of these models. The research paper [30] presented a new system that uses a Gated Recurrent Unit and Recurrent Neural Network for online handwritten mathematical equations. The author's proposed system consists of several components, including preprocessing, feature extraction, and recognition using the Gated Recurrent Unit (GRU) GRU-RNN network. The results show that the proposed system achieves a recognition accuracy of over 95% on the dataset, outperforming several state-of-the-art methods. C.T. Nguyen and H.T. Nguyen [31] proposed a novel approach to improve handwritten mathematical equation recognition by using a temporal classification constraint. The proposed approach uses a two-stage pipeline consisting of a segmentation stage and a recognition stage. The segmentation stage involves segmenting the input image into individual symbols, while the recognition stage involves recognizing the symbols using two neural networks convolution neural network followed by recurrent neural network with a temporal classification constraint. According to the results, the proposed model achieved greater results than other models in terms of recognition accuracy achieving an accuracy of 97% on the dataset. In the research [32], proposes a new neural network architecture for handwritten mathematical equation recognition. The author proposed a system called R-GRU, that system relies on a GRU network that has been regularized using a combination of dropout and L2 regularization techniques. The authors experiment with different dropout rates and regularization strengths to find the optimal combination for improving the recognition performance.

III. MATHVISION MODEL

A. PSEUDO ALGORITHM OF MODEL

Input: Image of a mathematical equation

Output: Audio description of the mathematical equation

It consists of four main modules equation detection and classification, encoder, decoder, and TTS module. YOLOv7 draws the bounding box around mathematical equations and classifies them into one of four categories integration, differentiation, limits, and trigonometry. Encoder is used to extract the feature map from the mathematical equation image. The decoder is responsible for generating natural language sentences. The attention mechanism is also used so that the decoder can focus on the relevant part of the mathematical equation image. The text-to-speech module takes natural language sentences generated by LSTM as input and converts them to voice. The integration of YOLOv7 DenseNet (CNN), and an attention-based LSTM in this research paper offers a comprehensive approach to convert mathematical equations into voice. YOLOv7 is utilized for accurate detection and categorization of equations, while a fine-tuned DenseNet extracts detailed features. The attention-based LSTM generates natural language descriptions, utilizing an attention mechanism for improved contextual relevance. This combined methodology ensures a robust and effective process, encompassing accurate detection, intricate feature extraction, and coherent natural language generation, thereby enabling a thorough conversion of mathematical content into voice. In the following subsections, each of these modules is discussed in detail.

1. Equation Detection and Classification using YOLOv7

Apply YOLOv7 to detect and classify potential equation regions in the image. a. Identify bounding boxes around potential equation areas. b. Classify each region as an equation or non-equation.

2. Feature Extraction using Fine-tuned DenseNet

For each detected equation region: a. Extract the sub-image from the original image corresponding to the bounding box. b. Pass the extracted equation image through the fine-tuned DenseNet model. c. Obtain the extracted feature representation for the equation image

3. Natural Language Generation using Attention-based LSTM

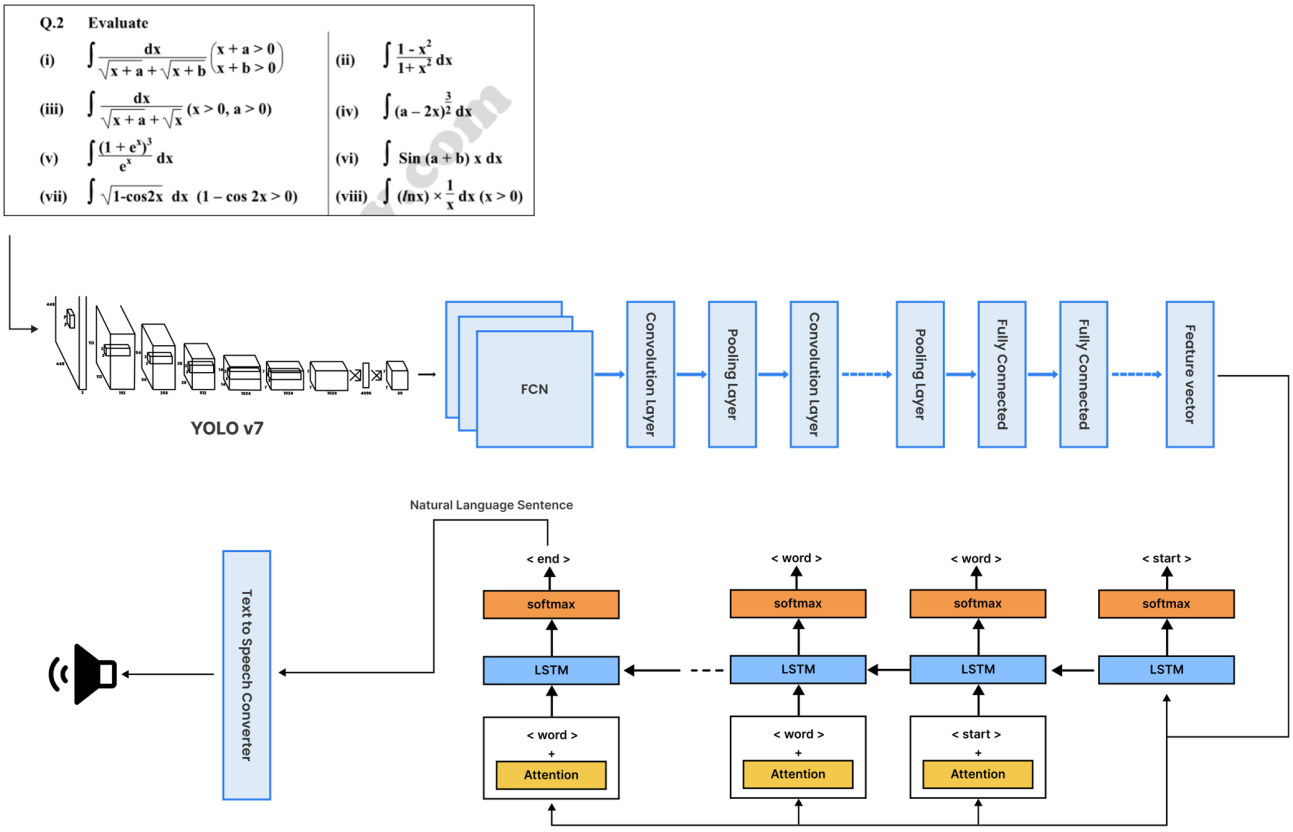


FIGURE 1. Overview of the proposed MathVision model with its modules. The model uses Yolov7 followed by CNN which is followed by LSTM and the last module is text to speech converter.

Combine the extracted feature representations from all detected equations. Initialize an attention-based LSTM model with the combined feature representation. Iteratively feed the LSTM model with attention mechanisms to generate a natural language sentence describing the mathematical equation.

4. Text-to-Speech Conversion using Python Library

Convert the generated natural language sentence into a suitable format for the text-to-speech (TTS) library. Utilize the Python TTS library to synthesize the audio representation of the mathematical equation.

B. MODEL OVERVIEW

MathVision model takes an image of a mathematical equation as an input and converts it to voice which gives a better understanding of mathematical equations for visually impaired people. The proposed MathVision model is illustrated in Fig. 1.

In the first step, we labeled our dataset. This labeled dataset contains four classes of mathematical equation limits, trigonometric functions, differentiation, and integration. Then we train the YOLOv7 on the mathematical equation dataset. YOLOv7 gives us state-of-the-art results. Fig. 3 shows the equation detection and classification with accuracy.

C. MathVision: EQUATION DETECTOR AND CLASSIFIER

Detection of real-time objects is an important task of computer vision. Object detection deals with the identification and location of objects within an image. It focuses on locating a region of interest inside an image and classifying this region in the manner of a standard image classifier. A rectangular box around the image shows that the object in the image is detected. Due to its speed and accuracy, YOLO (you only look once) is a widely used object detection algorithm. To process an image, YOLO uses a fully convolution neural network. It predicts bounding boxes and class probabilities simultaneously. As compared to Faster RCNN YOLO uses a single fully connected layer to carry out all of its predictions. The YOLO algorithm uses a straightforward deep convolutional neural network to identify objects in an input image. There are various versions of the YOLO model like YOLOv2, YOLOv3, YOLOv4, YOLOv5, YOLOv6. We used YOLOv7 which uses anchor boxes. To detect objects of various shapes, anchor boxes, a collection of preconfigured boxes with various aspect ratios, are employed. YOLOv7 uses nine anchor boxes, and it uses a new loss function called focal loss as compared to other models which use cross entropy. It also processes the image in a higher resolution of 608 by 608 pixel as compared to YOLOv3 which processes images at a resolution of 416 by 416. Speed and accuracy are the other

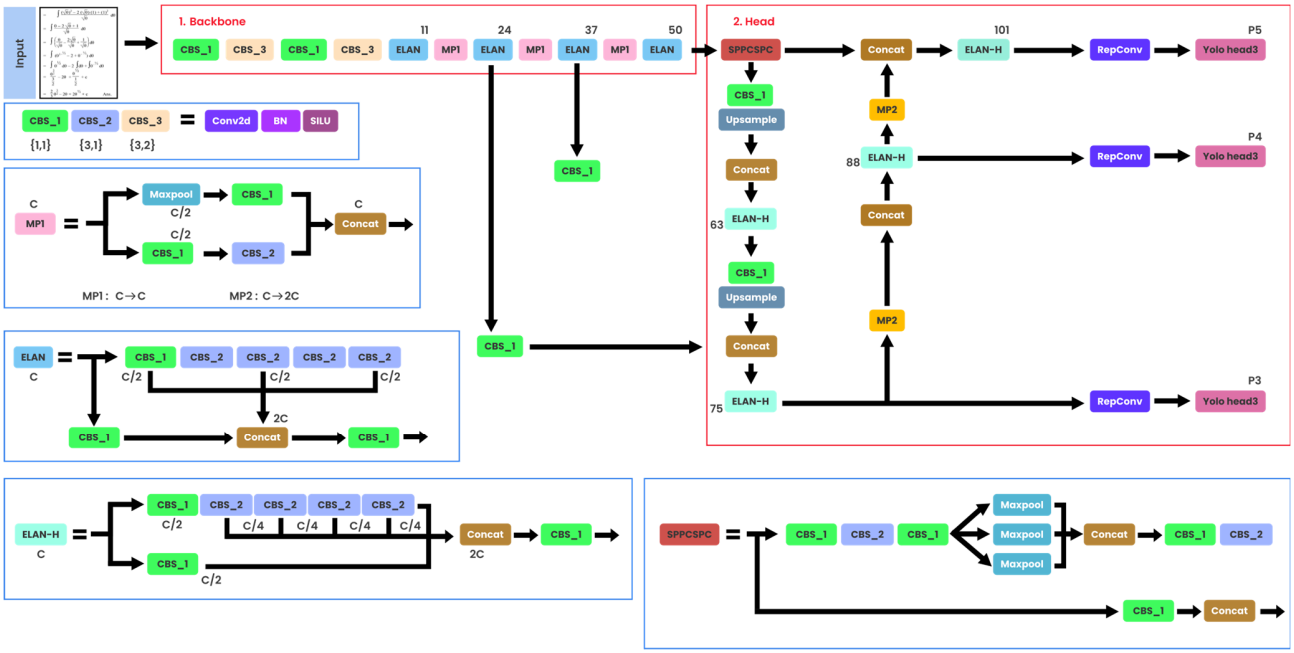


FIGURE 2. The network architecture diagram of YOLOv7 contains four general modules input terminal, backbone, head, prediction, and five basic components CBS, MP, ELAN, ELAN-H, and SPPCSPC.

main reason that inspires us to use YOLOv7 for equation detection and classification. The Network Architecture of YOLOv7 is shown in Fig. 2.

D. MathVision: Encoder

As the encoder, we employ the densely connected convolutional network (DenseNet) [12]. By ensuring maximal information flow between network layers, the DenseNet is Fully Convolution Neural Network (FCN) that connects all networks in a feed-forward fashion and strengthens feature propagation and reuse. Since mathematical equation images typically come in a variety of sizes, the FCN's ability to handle images of any size makes it suitable for feature extraction and equation recognition. A grayscale image is given as input to the encoder, having a size $1 \times H \times W$, where H *and* W are the height and width of an image. The encoder returns a $K \times H/\eta \times W/\eta$ matrix, which represents the channel number and η is down sampling factor. $E(Y) = [e_1, e_2, \ldots, e_L]$, represents encoding where and $e_i \in \mathbb{R}^K$. The local region of the image corresponds to each element $E(Y)$. We have implemented the encoder by setting it to 684 and 16 here 684 represents channel number and 16 represents down sampling factor. The architecture of a fully convolution neural network is shown in Fig. 4.

E. MathVision: Long short-term memory

The network structure of LSTM is built on top of RNN. LSTM memory cell structure is illustrated in Fig. 5.

A cell state is added to the network to deal with the problem of long-term sequence, which establishes whether past and present data can be joined via a gating mechanism, resolving the "gradient vanishing" and "gradient explosion" issues with the RNN. Information in the network is controlled through the use of three gates. How much of the previous state can be saved is determined by the forget gate. The input gate decides whether to update the LSTM's information using the current input. Which elements of the current cell state need to be transferred to the subsequent layer for iteration are determined by the output gate. The updating of a cell state can be broken down into the following steps.

1-Decide what unused data is removed from the previous time step's state;

2-The information that is valid and can be added to the state cell at the current time step should be extracted.

3- The current time stamp of the state unit is calculated

4- The current time stamp of the state unit is calculated

In this paper we consider LSTM [14] to implement a decoder that creates a sentence by creating a single word at a time based upon context vector $\hat{z}_t$hidden state h_t and the word that was formed by previous processing y_{t-1}. Eq. 1 is used to generate word at time step t:

$$p\left(y_t \mid y_1, y_2, \ldots, y_{t-1}, x\right) = f\left(y_{t-1}, h_t, \hat{z}_t\right) \tag{1}$$

Multi-Layered Perceptron (MLP) is denoted by f is enlarged in Eq (5) and x stands for the inputted mathematical equation image. Eq. 2 is used to calculate the hidden state of LSTM.

$$\begin{aligned} i_t &= \sigma\left(W_{yt} E y_{t-1} + U_{hi} h_{t-1} + V_{zt}\, \hat{z}_t\right) \\ f_t &= \sigma\left(W_{yf} E y_{t-1} + U_{hf} h_{t-1} + V_{zf}\, \hat{z}_t\right) \\ o_t &= \sigma\left(W_{yo} E y_{t-1} + U_{ho} h_{t-1} + V_{zo}\, \hat{z}_t\right) \\ g_t &= \tanh\left(\left(W_{yc} E y_{t-1} + U_{hc} h_{t-1} + V_{zc}\, \hat{z}_t\right)\right) \\ c_t &= f_t\, {}^{*}\, c_{t-1} + i_t\, {}^{*}\, g_t \end{aligned}$$

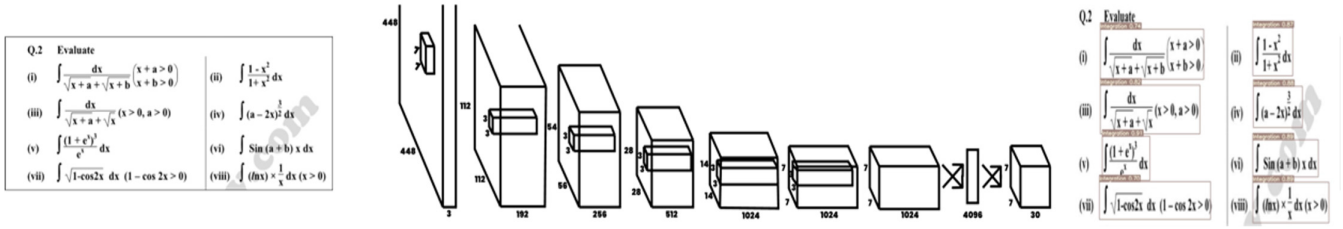


FIGURE 3. Sample classification results of yolov7.The red bounding box represents equation detection and the upper corner of the image class of mathematical equation is detected by yolov7 with percentage accuracy.

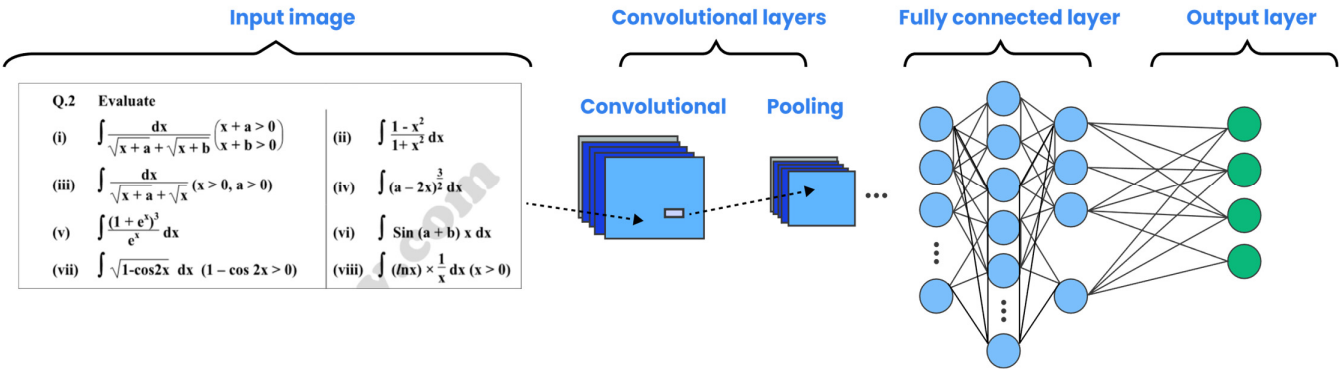


FIGURE 4. The architecture of a fully connected convolution network.

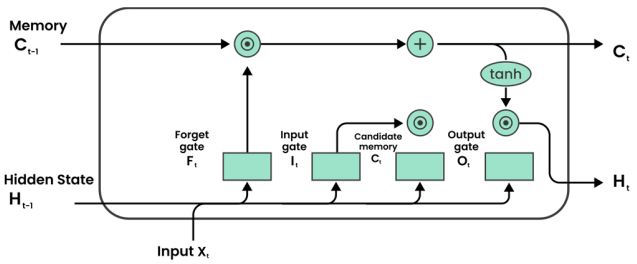


FIGURE 5. Long short-term memory cell structure. X represents the input cell of the LSTM memory cell. The light green circle is the operator for calculating forget gate, input gate, and output gate. The calculation formula for generating forget, input and output gate is denoted by a rectangle.

$$h_t = o_t \ ^* \ \tanh(c_t) \tag{2}$$

i_t, f_t, c_t, o_t and h_t are the LSTM states like input, forget, memory, output, and hidden state respectively. The vector $\hat{z}_t$ is a context vector that stores the visual data for a specific area of an image. The relevant area of the image at time step t is dynamically represented by the context vector $\hat{z}_t$ in Eq 2.

F. ATTENTION MECHANISM

We are considering a soft attention mechanism in which weights α_{ti} of each annotation vector are computed based on the previous LSTM hidden state h_{t-1}. Jointly trained attention is parametrized as MLP in Eq. 3.

$$e_{ti} = v_a^t \tanh(W_a h_t + U_a \partial_i)$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})} \tag{3}$$

Let say the dimension of attention is represented by n' then $v_a \in R^{n' \times n}$, $U_a \in R^{n' \times D}$ [14]. The context vector $\hat{z}_t$is computed using Eq. 4 after the calculation of weightsα_{ti}

$$\hat{z}_t = \sum_{i=1}^{l} \alpha_{ti} \partial_i \tag{4}$$

This weight α_{ti} helps the decoder determine which area of the input image it focuses on to produce the subsequent predicted word, and it then gives that area more weight in the related annotation vectors ∂_i m and n represent LSTM and embedding dimensions. The expression $E \in R^{m \times k}$ [14] denotes the embedding matrix. $*$ represents element-wise multiplication σdenotes sigmoid activation function.

Finally, using the context vector $\hat{z}_t$, current LSTM hidden state h_t, and previous predicted word y_{t-1}, the predicted word probability at time t is calculated in Eq. 5:

$$p(y_t \mid y_1, y_2, \ldots, y_{t-1}, x) = g\left(W_o\left(Ey_{t-1} + W_h h_t + W_z \hat{z}_t\right)\right) \tag{5}$$

Across all the vocabulary words g represents softmax activation function.

E, $W_o R^{K \times n}$, $W_h \in R^{m \times n}$ *and* $W_c \in R^{m \times D}$ are known as parameters that are initialized randomly.

An average of the annotation vectors fed through two different MLPs $(finit, c, finit, h)$ predicts the initial memory state c_o and hidden state h_o of the LSTM. Eq. 6 represents the

initial state and hidden state of LSTM.

$$c_0 = finit,\ c(\frac{1}{L}\sum_{i=1}^{L}\partial_i)$$
$$h_0 = finite,\ h(\frac{1}{L}\sum_{i=1}^{L}\partial_i) \quad (6)$$

G. TEXT-TO-SPEECH MODULE

We employed text to text-to-speech system proposed by Girish.et.al. Proposed text-to-speech system will take input in the form of a natural language sentence generated by the LSTM model and convert it to voice. This TTS using natural language processing analyzes and processes the text and then converts this processed text into speech using digital signal processing technology. A platform-independent pyttsx3 python library is used for converting text into speech. The main reason for using this library is that it works offline. The figure given below illustrates the system design.

H. IMPLEMENTATION DETAILS

To train the model, we take into account 9.6K images and their related natural language descriptions. For example, an image named INTG1.png was given a natural language description of ''integration of cos x with respect to dx'' Similarly all the 9.6K images were given a natural language description. In the first phase of the implementation, we trained YOLO v7 for the detection and classification of mathematical equations. In the next phase, we employed DenseNet as an encoder which is a fully connected convolution network. We trained proposed model with a batch size of 100 with 50 epochs. Stochastic gradient descent is used with a constant learning rate of 0.0001 and weight decay of 0.0001. Except for FCN, all other weights are initialized randomly. For the embedding, we take 512 dimensions, while the size of the LSTM memory is 1024. After each convolution layer, we added a dropout layer to LSTM and set it to 0.5. The output of LSTM, which is a natural language sentence, is evaluated using the widely used Natural Language Processing (NLP) metric BLEU score. In the last phase of the implementation, we use the pyttsx3 python library which is used for text-to-speech conversion. Proposed model outputs in the form of voice therefore using this library natural language sentence is converted into voice. This voice output is used by visually impaired people which helps them a lot for understanding mathematical equations. The proposed model generates the voice of mathematical equation in 4.438s.

IV. DATASET AND EVALUATION METRICS

A. DATASET

We created the dataset as there is no publicly available dataset along with their natural language descriptions. The link for datasethttps://github.com/MohammadAwaisAhmad/MathVision-DLMODEL.We take images from Punjab textbook board higher secondary school book. We limit ourselves to four classes of mathematical equations limits, trigonometric functions, differentiation, and integration. Our dataset consists of 12K images. For the dataset, sets of predefined variables, operators, and constants—such as x, y, and z—as well as sets of their related natural language descriptions are produced. The mathematical equation is automatically generated as an image and a textual explanation in the text format using a Python code that randomly chooses a function, variable, operator, and constant from the relevant predefined sets. The dataset is split into train tests with ratios of 80% and 20% respectively.

TABLE 1. Performance comparison between MED [14] with proposed math vision model.

	Performance Comparison				
Model	B1	B2	B3	B4	Rouge
MED[14]	0.975	0.959	0.936	0.907	9.197
MathVision	0.983	0.965	0.943	0.926	9.249

B. EVALUATION METRICS

In this research, we evaluated the performance of proposed MathVision model till the LSTM module using Bilingual Evulation Understudy (BLEU) score used in [33] alongside Recall-Oriented Understudy for Gisting Evaluation (ROUGE). These metrics are frequently employed in NLP and image captioning applications. BLEU metric essentially assesses how closely an LSTM-generated sentence resembles a collection of reference truth sentences created by humans. The LSTM-generated sentence (text) is more comparable to the Reference truth sentence when all of these measures have higher values (text). On the other hand, ROUGE evaluates generated summaries by comparing them to human-generated reference summaries. It involves collecting reference summaries, tokenizing both sets into n-grams, and calculating the overlap of these n-grams to derive precision, recall, and F1 scores. Higher ROUGE scores indicate better summary quality due to greater overlap with references. We compare proposed LSTM module results with [14], proposed model achieves significant improvement in term of BLEU score and ROUGE. We used four variants of BLEU. These are BLEU1, BLEU2, BLEU3, and BLEU4. We trained the MED model on our dataset and then BLEU and ROUGE scores are re-evaluated. The Table 1 shows the comparison of proposed model with MED [14].

We compared MathVision Model with MED based on number of computations involved. The MathVision execution proceeds in timesteps, with a new symbol of the input sequence processed at each timestep. MathVision is computationally more intensive than MED as MathVision performs O $4(n^2)$ Computations per timestamp and MED performs O $3(n^2)$ Computations per timestamp. Moreover, MED requires three matrix multiplications per time step, while MathVision requires four, leading to a more computationally efficient model. Fig. 6 shows the efficiency of MathVision Model as compared to MED [14]. The sentence generated by the both MathVision and MED is compared against the ground truth sentence written by humans. MathVision generated almost

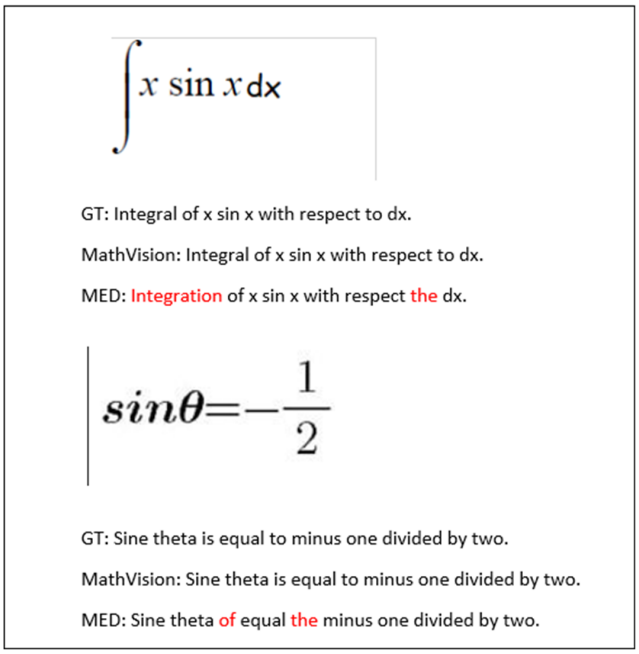


FIGURE 6. Comparison of sample results produced by MathVision Model and MED [14]. Red colored text indicates wrongly generated text.

same sentence written by humans while MED misses some of text which is highlighted in red text. This shows MathVision results are closer to the ground truth as compared to MED.

V. RESULTS

We have included the screenshot of some of the sample results, which are the output of our MathVision model shown in Fig. 7. The voice output shows that visually impaired people can easily understand mathematical equations. We tested the effectiveness of our model by giving some random equation images. Out of 10 equations, nine were converted successfully to the speech. Then, these nine equations were written successfully by visually impaired students.

VI. CONTRIBUTIONS

To check the effectiveness of proposed MathVision model we conducted real-world experiments by visiting a high school. For this purpose, we have taken images from the class IX mathematics textbook. Our set consists of 100 mathematical equation images which contain four categories' limits, trigonometry, integration, and differentiation. Our MathVision model has successfully generated the audio of these equations. Then, these audios are provided to visually impaired students, and they are asked to write mathematical equations. Fifteen students participated in this test. If a student writes an equation correctly then its answer is correct and if a student writes an equation wrong, then its answer is incorrect. Among 100 equations, students have written 95 equations correctly within a given time range by listening to the audio generated by proposed MathVision model. For the remaining five equations, proposed MathVision model

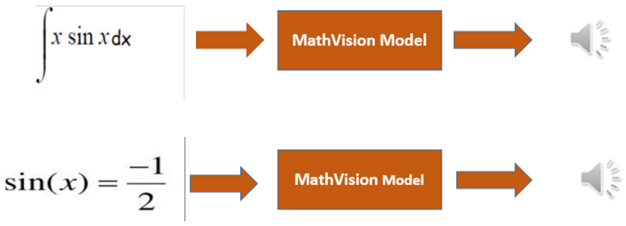


FIGURE 7. Sample audios generated by our MathVision model.

generated incorrect audio due to the presence of other equations like binomial theorem and different geometrical shapes. Since the audio generated by the proposed MathVision model is not correct students write incorrect equations by listening wrong audio. This test shows that proposed model is highly effective in helping visually impaired people to understand mathematical equations.

VII. FUTURE WORK

The proposed encoder module can be used in bioinformatics such as to analyze the biological datasets (e.g., gene expression) to find hidden patterns and drug targets. It can also be used to analyze high-throughput imaging data (e.g., cell biology) for faster diagnoses and deeper understanding. Integration of proposed model with existing bioinformatics tools for wider applicability. In image processing our decoder when integrated with the convolutional neural network can be used to solve image captioning problems. It can also be used in other image-intensive fields like microscopy, and remote sensing and can effectively process sequential image data. In the future we are planning to improve our work by using some deep neural networks used in image processing and bioinformatics.

VIII. CONCLUSION

In this paper, a novel model named MathVision is introduced which generates the voice of mathematical equations. This voice is helpful for visually impaired students to understand mathematical equations. We have created our dataset due to the unavailability of mathematical equations datasets with their natural language description. Real-word experiment proves that visually impaired students are comfortable to write mathematical equations accurately by listening to the audio generated as output by proposed MathVision model. This experiment shows the effectiveness of the model and its tremendous impact on the visually impaired community working in the education sector. Moreover, visually impaired students might face technological barriers in adopting the MathVision Model such as output compatibility with screen readers and braille displays, achieving accurate image interpretation of diverse mathematical notations, and providing real-time processing capabilities are significant challenges. Additionally, seamless integration with existing educational tools and adherence to ethical guidelines for privacy and bias are crucial for successful adoption.

ACKNOWLEDGMENT

This research was supported by Princess Nourah bint Abdulrahman University and Researchers Supporting Project number (PNURSP2025R346).

The authors would also like to thanks AIDA Lab CCIS Prince Sultan University, Riyadh Saudi Arabia for their support.

REFERENCES

[1] World Health Org. (2023). *Blindness and Vision Impairment*. [Online]. Available: https://www.who.int/news-room/factsheets/detail/blindness-and-visual-impairment

[2] M. S. Oyebanji and U. S. Idiong, "Challenges of teaching mathematics to students with visual impairment," *Malikussaleh J. Math. Learn. (MJML)*, vol. 4, no. 1, p. 1, May 2021, doi: 10.29103/mjml.v4i1.2538.

[3] J. Ganesan, A. T. Azar, S. Alsenan, N. A. Kamal, B. Qureshi, and A. E. Hassanien, "Deep learning reader for visually impaired," *Electronics*, vol. 11, no. 20, p. 3335, Oct. 2022.

[4] D. AlSaeed, H. Alkhalifa, H. Alotaibi, R. Alshalan, N. Al-Mutlaq, S. Alshalan, H. T. Bintaleb, and A. M. AlSahow, "Accessibility evaluation of Saudi e-government systems for teachers: A visually impaired user's perspective," *Appl. Sci.*, vol. 10, no. 21, p. 7528, Oct. 2020.

[5] S. He, H. Luo, W. Jiang, X. Jiang, and H. Ding, "VGSG: Vision-guided semantic-group network for text-based person search," *IEEE Trans. Image Process.*, vol. 33, pp. 163–176, 2024, doi: 10.1109/TIP.2023.3337653.

[6] M. N. A. Wahab, A. S. A. Mohamed, A. S. A. Sukor, and O. C. Teng, "Text reader for visually impaired person," *J. Phys., Conf. Ser.*, vol. 1755, no. 1, Feb. 2021, Art. no. 012055, doi: 10.1088/1742-6596/1755/1/012055.

[7] J. B. Dheesha. (2022). *Teaching Mathematics To Students With Visual Impairment With Screen Reading Software*. [Online]. Available: https://www.researchgate.net/publication/363481219_Teaching_Mathematics_to_Students_with_Visual_Impairment_With_Screen_Reading_Software

[8] X. Shen, H. Jiang, D. Liu, K. Yang, F. Deng, J. C. S. Lui, J. Liu, S. Dustdar, and J. Luo, "PupilRec: Leveraging pupil morphology for recommending on smartphones," *IEEE Internet Things J.*, vol. 9, no. 17, pp. 15538–15553, Sep. 2022, doi: 10.1109/JIOT.2022.3181607.

[9] D. Alsaleh and S. Larabi-Marie-Sainte, "Arabic text classification using convolutional neural network and genetic algorithms," *IEEE Access*, vol. 9, pp. 91670–91685, 2021.

[10] J. Ding, X. Chen, P. Lu, Z. Yang, X. Li, and Y. Du, "DialogueINAB: An interaction neural network based on attitudes and behaviors of interlocutors for dialogue emotion recognition," *J. Supercomput.*, vol. 79, no. 18, pp. 20481–20514, Dec. 2023, doi: 10.1007/s11227-023-05439-1.

[11] J.-W. Wu, F. Yin, Y.-M. Zhang, X.-Y. Zhang, and C.-L. Liu, "Handwritten mathematical expression recognition via paired adversarial learning," *Int. J. Comput. Vis.*, vol. 128, nos. 10–11, pp. 2386–2401, Jan. 2020, doi: 10.1007/s11263-020-01291-5.

[12] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao. (2022). *YOLOv7: Trainable Bag-of-Freebies Sets New State-of-The-Art for Real-Time Object Detectors*. [Online]. Available: https://arxiv.org/abs/2207.02696

[13] Y. Yuan. (2022). *Syntax-Aware Network for Handwritten Mathematical Expression Recognition*. [Online]. Available: https://arxiv.org/abs/2203.01601

[14] A. Mondal and C. V. Jawahar, "Textual description for mathematical equations," 2020, *arXiv:2008.02980*.

[15] J. Xu, K. Guo, X. Zhang, and P. Z. H. Sun, "Left gaze bias between LHT and RHT: A recommendation strategy to mitigate human errors in left- and right-hand driving," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 10, pp. 4406–4417, Oct. 2023, doi: 10.1109/TIV.2023.3298481.

[16] W. Zhao, L. Gao, Z. Yan, S. Peng, L. Du, and Z. Zhang, "Handwritten mathematical expression recognition with bidirectionally trained transformer," 2021, *arXiv:2105.02412*.

[17] P. Tope, S. Ransubhe, M. A. Mughni, C. Shiralkar, and M. B. Ratnaparkhi, "Recognition of handwritten mathematical expression and using machine learning approach," *Int. Res. J. Eng. Technol.*, vol. 3, pp. 1–26, May 2021.

[18] D. Ogwok and E. M. Ehlers. (2020). *Detecting, Contextualizing and Computing Basic Mathematical Equations From Noisy Images Using Machine Learning*. [Online]. Available: https://dl.acm.org/doi/fullHtml/10.1145/3440840.3440855

[19] P. U. Patil and K. N. Khot, "Handwritten mathematical expression recognition and grading system," *Int. J. Eng. Appl. Sci. Technol.*, vol. 6, no. 3, pp. 1–16, Jul. 2021.

[20] R. Aggarwal, S. Pandey, A. K. Tiwari, and G. Harit, "Survey of mathematical expression recognition for printed and handwritten documents," *IETE Tech. Rev.*, vol. 39, no. 6, pp. 1245–1253, Dec. 2021, doi: 10.1080/02564602.2021.2008277.

[21] V. Kukreja, "A retrospective study on handwritten mathematical symbols and expressions: Classification and recognition," *Eng. Appl. Artif. Intell.*, vol. 103, Aug. 2021, Art. no. 104292, doi: 10.1016/j.engappai.2021.104292.

[22] Z. Li, L. Jin, S. Lai, and Y. Zhu, "Improving attention-based handwritten mathematical expression recognition with scale augmentation and drop attention," 2020, *arXiv:2007.10092*.

[23] B. H. Pong, L. T. Dot, Thang, M. Haong, and T. L. Le. (2020). *A Deep Learning Based System for Mathematical Equation Detection and Recognition in Document Images*. [Online]. Available: https://ieeexplore.ieee.org/document/9287693

[24] W. Zhang, Z. Bai, and Y. Zhu. (2019). *An Improved Approach Based on CNN-RNNs for Mathematical Expression Recognition*. [Online]. Available: https://dl.acm.org/doi/10.1145/3330393.3330410

[25] J. Dai, Y. Sun, G. Su, S. Ye, and Y. Sun. (2019). *Recognizing Offline Handwritten Mathematical Expressions Efficiently*. [Online]. Available: https://dl.acm.org/doi/10.1145/3306500.3306543

[26] J.-W. Wu, F. Yin, Y. Zhang, X.-Y. Zhang, and C. Liu, "Graph-to-graph: Towards accurate and interpretable online handwritten mathematical expression recognition," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 4, pp. 2925–2933, doi: 10.1609/AAAI.V35I4.16399.

[27] X. Bian, B. Qin, X. Xin, J. Li, X. Su, and Y. Wang, "Handwritten mathematical expression recognition via attention aggregation based bi-directional mutual learning," *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 1, pp. 113–121, Jun. 2022, doi: 10.1609/aaai.v36i1.19885.

[28] S. Shinde, A. Mulagirisamy, D. G. Bhalke, and L. Wadhwa, "Recognition of math expressions & symbols using machine learning," *Turkish Online J. Qualitative Inquiry*, vol. 12, no. 8, pp. 1–10, Oct. 2021.

[29] D. Zhelezniakov, V. Zaytsev, and O. Radyvonenko, "Online handwritten mathematical expression recognition and applications: A survey," *IEEE Access*, vol. 9, pp. 38352–38373, 2021, doi: 10.1109/ACCESS.2021.3063413.

[30] Z. Endong and L. Liu. (2021). *Design of Online Handwritten Mathematical Expression Recognition System Based on Gated Recurrent Unit Recurrent Neural Network*. [Online]. Available: https://ieeexplore.ieee.org/document/9551034

[31] C. T. Nguyen, H. T. Nguyen, K. Morizumi, and M. Nakagawa, "Temporalclassification constraint for improving handwritten mathematical expression recognition," in *Proc. Int. Conf. Document Anal. Recognit.*, 2021, pp. 113–125.

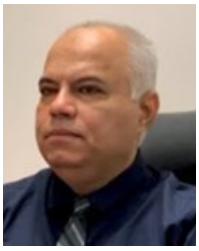
[32] A. Pal and K. P. Singh, "R-GRU: Regularized gated recurrent unit for handwritten mathematical expression recognition," *Multimedia Tools Appl.*, vol. 81, no. 22, pp. 31405–31419, Apr. 2022, doi: 10.1007/s11042-022-12889-x.

[33] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics*, 2001, p. 311.

MUHAMMAD AWAIS AHMAD received the B.S. degree in CS from the Department of Computer Science, Lahore Garrison University (LGU), Lahore, Pakistan, in 2019, and the M.S. degree in CS from the Department of Computer Science, UET Lahore, Pakistan, in 2023. His research interests include deep learning, blockchain security, and decentralized finance.

TAUQIR AHMED received the Doctor of Philosophy (Ph.D.) degree in computer science from the University of Engineering and Technology (UET), Lahore, in 2012. He has been an Associate Professor with the Department of Computer Science and Engineering, UET Lahore, since 1999, performing duties, like teaching and research.

MUHAMMAD ASLAM received the Ph.D. degree in computer sciences from CINVESTAV-IPN, Mexico, in 2005, under the Cultural Exchange Scholarships between Pakistan and Mexico. He has more than 15 years of experience in software architecture design, team leading, team building, and software projects. He has 14 years of experience in research and development and teaching at the postgraduate level (supervising Ph.D. and M.Sc. thesis). He has supervised six Ph.D. students and more than 50 master's thesis and final-year projects. He has published his research findings in several well-reputed impact factor journals, like Springer and IEEE. His research interests include artificial intelligence, distributed intelligence, knowledge-based systems, expert systems, intelligent agents, human–computer interaction, machine learning, computer-supported cooperative work, cooperative writing and authoring, cooperative learning, and distributed computing. He won the Merit Scholarship from the Board of Intermediate and Secondary Education, Sargodha Division, Pakistan, from 1984 to 1986. He won the Silver Medal from the Faculty of Agricultural Engineering, University of Agricultural, Faisalabad, Pakistan, from 1987 to 1991. He was also awarded the Cultural Exchange Scholarship between Pakistan and Mexico, from 2000 to 2004, for the Ph.D. studies.



AMJAD REHMAN (Senior Member, IEEE) received the Ph.D. degree from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2010, with a focus on information security using image processing techniques. He is currently an Associate Professor with CCIS, Prince Sultan University, Riyadh, Saudi Arabia. He is also a PI of several projects and completed projects funded by MoHE Malaysia, Saudi Arabia. His research interests include bioinformatics, the IoT, information security, and pattern recognition. He received the Rector Award for the 2010 Best Student from UTM Malaysia.

FATEN S. ALAMRI received the Ph.D. degree in system modeling and analysis in statistics from Virginia Commonwealth University, USA, in 2020. Her Ph.D. research included Bayesian dose-response modeling, experimental design, and nonparametric modeling. She is currently an Assistant Professor with the Department of Mathematical Sciences, College of Science, Princess Nourah bint Abdul Rahman University. Her research interests include spatial area, environmental statistics, and brain imaging.

SAEED ALI BAHAJ received the Ph.D. degree from Pune University, India, in 2006. He is an Associate Professor with the MIS Department, COBA Prince Sattam bin Abdul-Aziz University KSA. He has published above 50 papers in highly reputed journals and also won funded projects. His main research interests include information management, forecasting, information engineering, and information security.

TANZILA SABA (Senior Member, IEEE) received the Ph.D. degree in document information security and management from the Faculty of Computing, Universiti Teknologi Malaysia (UTM), Malaysia, in 2012. Currently, she is a Full Professor with the College of Computer and Information Sciences, Prince Sultan University (PSU), Riyadh, Saudi Arabia, and also the AIDA Laboratory Leader. She has published over 300 publications in high-ranked journals. Her research interests include bioinformatics, data mining, and classification using AI models. She won the Best Student Award from the Faculty of Computing, UTM, in 2012. She was awarded the Best Researcher of the Year Award at PSU, from 2013 to 2016. Due to her excellent research achievement, she is included in Marquis Who's Who (S & T) 2012. She is an editor of several reputed journals and on a panel of TPC of international conferences.