Minor Project Report

on

Mathematical Expression Localization in Handwritten Document

by

Tarang Ghetia
[20BCP300]
Mahir Mehta
[20BCP122]

Under the Guidance of
Dr. Shilpa Pandey
Assistant Professor

Submitted to

Department of Computer Science and Engineering

School of Technology

Pandit Deendayal Energy University

2023

Minor Project Report

on

Mathematical Expression Localization in Handwritten Document

by

Tarang Ghetia
[20BCP300]
Mahir Mehta
[20BCP122]

Under the Guidance of
Dr. Shilpa Pandey
Assistant Professor

Submitted to

Department of Computer Science and Engineering

School of Technology

Pandit Deendayal Energy University

2023

CERTIFICATE

This is to certify that the seminar report entitled “Mathematical Expression Localization in Handwritten Document,” submitted by Tarang Ghetia and Mahir Mehta, has been conducted under the supervision of Dr. Shilpa Pandey, Assistant Professor, and is hereby approved for the partial fulfillment of the requirements for the award of the degree of Bachelor of Engineering in the Department of Computer Science & Engineering at Pandit Deendayal Energy University, Gandhinagar. This work is original and has not been submitted to any other institution for the award of any degree.

Sign:
Dr. Shilpa Pandey
Assistant Professor
Computer Science & Engineering
School of Technology
Pandit Deendayal Energy University

Sign:
Dr. Hargeet Kaur
Assistant Professor
Computer Science & Engineering
School of Technology
Pandit Deendayal Energy University

ACKNOWLEDGEMENT

It is with immense satisfaction and glee that I extend my heartfelt gratitude and appreciation to Pandit Deendayal Energy University in Gandhinagar for their invaluable support and resources that have made this journey possible. I would like to extend my sincerest thanks to my mentor, Dr. Shilpa Pandey of the Computer Scienceand Engineering department at SOT, PDEU. Her insights, suggestions, and constant encouragement have been instrumental in the successful development of this project. Moreover, her interest and confidence in me, both as a person and as a professional, have been a great influence that enabled meto achieve this feat. My gratitude goes out to the staff of the Department of SOT at Pandit DeendayalEnergy University for their relentless guidance and cooperation.

The completion of this project has been a journey of collective effort, and I would like to express my deepest gratitude to all the individuals who've been part of this journey. Finally, it would be remiss of me to overlook the ultimate cause of this journey, without Whose blessings this would not have been possible. Thus, I express my heartfelt thanks to the Prime Mover for His continual showers of blessings.

Mahir Mehta	Tarang Ghetia
20BCP122	20BCP300

DECLARATION

I hereby declare that the seminar report entitled “Mathematical Expression Localization Using Handwritten Document” is the result of my own work and has been written by me. This report has not utilized any language model or natural language processing artificial intelligence tools for the creation or generation of content, including the literature survey.

The use of any such artificial intelligence-based tools was strictly confined to the polishing of content, spellchecking, and grammar correction after the initial draft of the report was completed. No part of this report has been directly sourced from the output of such tools for the final submission.

This declaration is to affirm that the work presented in this report is genuinely conducted by me and to the best of my knowledge, it is original.

Tarang Ghetia
20BCP300
Computer Science & Engineering
School of Technology
Pandit Deendayal Energy University
Gandhinagar

Mahir Mehta
20BCP122
Computer Science & Engineering
School of Technology
Pandit Deendayal Energy University
Gandhinagar

Date: 22nd Nov, 2023
Place: Gandhinagar

List of Tools Used for the Report with Purpose:

- **ChatGPT: Correcting Grammar.**
- **Anaconda: For coding.**
- **Google collab: For implementing the code**
- **Google Drive: Storage**

ABSTRACT

Deep learning, mainly using convolutional neural networks (CNNs), has revolutionized computer vision by enabling machines to learn and interpret images independently. CNNs are highly proficient in various image processing tasks, including object recognition, image segmentation, and optical character recognition (OCR). In this project, we leverage deep learning to identify and locate mathematical expressions in digital images. Mathematical expressions are crucial in scientific research, education, and technical documents, however, recognizing and understanding them can be challenging due to the diverse symbols, superscripts, subscripts, and mathematical notations. Additionally, accurately pinpointing mathematical equations in images is complicated by variations in font styles, sizes, and image resolutions. Precisely locating these mathematical expressions is a crucial step in automating their recognition.

Accurately locating mathematical expressions is crucial in various fields, including education, science, and document analysis. Automated mathematical formula recognition using OCR systems with precise mathematical expression localization capabilities makes it easier to understand complex mathematical expressions. Different techniques have been employed to address the issue of localizing mathematical expressions, such as edge detection, template matching, and deep learning-based approaches. While edge detection and template matching methods use predetermined templates or extract edges from the input image to identify mathematical symbols, deep learning-based techniques utilize convolutional neural networks (CNNs) to extract features from the input image and generate pixel-level predictions for the presence of mathematical symbols.

In this project, our proposed approach for localizing mathematical expressions utilizes the U-Net fully convolutional network (FCN) architecture. The U-Net FCN architecture is a deep learning-based technique that has exhibited excellent performance in various image segmentation tasks. To generate pixel-level predictions for the presence of mathematical symbols in the input image, we utilize the U-Net architecture in our approach. The successful implementation of the U-Net FCN in our method demonstrates its potential for image segmentation tasks, and further research could explore its suitability for other areas of image processing and computer vision.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Nomenclature

Abbreviations

OCR	Optical Character Recognition
FCNs	Fully Convolutional Networks
CNNs	Convolutional Neuron Networks
U-Net FCN	Unit-Net Fully Convolutional Network

CHAPTER 1

INTRODUCTION

1.1 Introduction

Deep learning-based models have shown impressive performance in a number of images processing tasks, including object detection, imagine classification, and optical character recognition (OCR) [1]. Fully convolutional networks (FCNs) in particular have been utilized extensively for segmentation tasks, where the objective is to categorize each pixel in an image [2].Due to its capability to extract high-level characteristics from the input picture and produce predictions at the pixel level, U-Net architecture among the FCNs has demonstrated outstanding performance in a multiple image segmentation task.

Popular image segmentation architecture U-Net FCN has been demonstrated to deliver cutting-edge performance in a number of applications. One of U-Net FCN's main benefits is its capacity to extract fine details from the input image while preserving spatial information. Skip connections are used to do this, allowing the network to integrate low-level encoder capabilities with high-level decoder characteristics [3]. This capability is particularly helpful for precisely identifying tiny, complicated signs and symbols that are frequently used in mathematical expressions in the context of localization. A versatile option for the localization of mathematical

expressions in both high- and low-resolution pictures, the U-Net architecture is also extremely adjustable and can be simply adjusted to accommodate different input image resolutions [4].

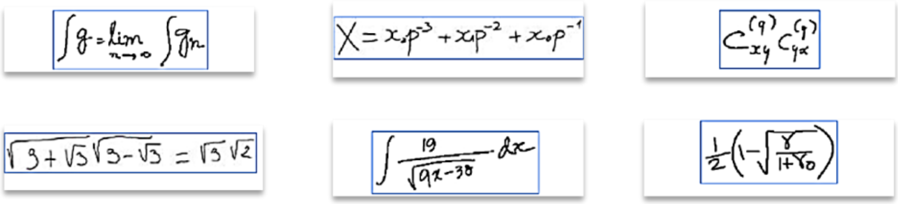
The localization of mathematical expressions using U-Net FCN architecture has important ramifications for analyzing documents and recognition. In the fields of science, engineering, and education, mathematical expressions are often utilized, and their precise localization and detectioncan speed up numerous procedures and improve the effectiveness of various applications. The adoption of deep learning-based approaches, such as U-Net FCN, can enable quicker and more accurate localization of mathematical expressions, improving detection and analysis due to the growing number of digital documents and the requirement for automated document processing.

The relevance of the challenge of localizing mathematical expressions in digital images isdiscussed in this work from the perspective of Web 3.0, and the U-Net FCN architecture is then introduced, along with an explanation of how it might be used to the localization of mathematicalequations in digital images. Finally, we describe the experimental findings made possible by our suggested methodology and talk about its potential uses in a range of industries, including research,education, and analysis of documents.

The suggested method offers a reliable and effective method for localizing mathematical expressions, which can greatly increase the precision and efficiency of document examination and recognition systems. The promise of deep learning-based techniques in processing images and computer vision applications is illustrated by the usage of U-Net FCN in this strategy.

1.2 Mathematical Notations

For the purpose of expressing mathematical ideas and actions, mathematical notations actas a symbolic language. the operators, brackets, Numbers, parameters, and unique characters are just a few of the symbols that they include. In mathematical formulas notations are essential because they enable the clear and accurate description of mathematical ideas.



$\int g = \lim_{n \to \infty} \int g_n$
$X = x_0 p^{-3} + x_1 p^{-2} + x_0 p^{-1}$
$C_{xy}^{(q)} C_{yx}^{(q)}$
$\sqrt{3+\sqrt{3}}\sqrt{3-\sqrt{3}} = \sqrt{3}\sqrt{2}$
$\int \frac{19}{\sqrt{9x-30}} dx$
$\frac{1}{2}\left(1-\sqrt{\frac{\gamma}{1+\gamma_0}}\right)$

Fig 1.1. Mathematical Notations

1.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are made specifically for image identification and handling tasks. Due to their capacity to analyse photos utilising subset filters that can recognise abstract structures and patterns in photographs, they have experienced a major increase inpopularity in recent years. Convolutional layers, which apply filters to the input picture and learn

to recognise relevant representations at various spatial positions, are a crucial component of CNNs. Pooling layers, which down-sample the feature maps to minimise the computational effort and extract just the most prominent features, are then used to further process these learnt features. Thefeature maps that are produced may be used to a number of tasks, including face recognition, objectidentification, and picture categorization.

The flexibility of CNNs to handle photos of various sizes and resolutions is another crucial aspect. CNNs can recognise a dog in a picture regardless of its location or orientation since they automatically tackle the translation invariance problem. Since pictures may change in size, inclination, and form, CNNs are well suited for applications like object detection and image categorization. There are several methods to construct CNNs, but popular structures include LeNet-5, AlexNet, and VGG-16.

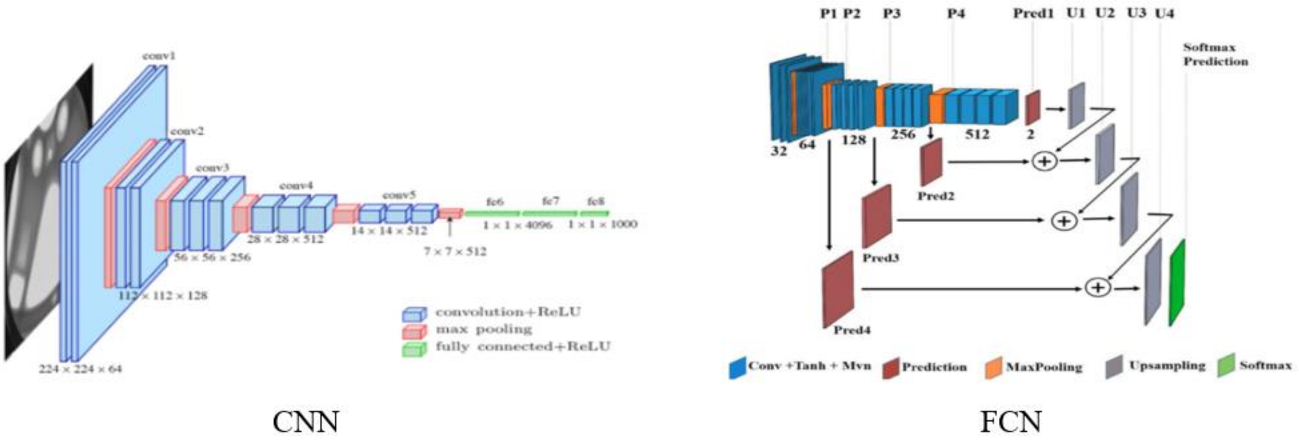
1.4 FCNs as Regressors

In contrast to discrete predictions, fully-convolutional networks (FCNs) enable image regression problems where the result is a continuous value or group of values. These networks have grown in popularity recently since they can categorize and segment pictures without the needto modify the model for each distinct purpose.

In semantic segmentation, which aims to give label to each pixel in an image specifyingwhich item or class it belongs to, FCNs are used as a regressor. A SoftMax function that is appliedto each pixel, computes the distribution of probabilities across all conceivable classes can be used

to do this. The binary segmentations, which may be used to distinguish various objects or areas from the backdrop of the picture, are created by thresholding the probability map that results.

FCNs as a regressor can be created in different ways, but one of the primary architectures is the U-Net architecture. With a succession of convolutional and pooling layers for feature extraction, it is intended to take advantage of the highly correlated nature of picture pixel information. Then, on the upsampled picture, the filter outputs from the same level of encoding are concatenated using the feature maps to achieve upscaling using deconvolutions.



CNN

FCN

Fig 1.2. CNN vs FCN

1.5 Saliency maps and Binary Masks

Saliency maps and binary masks are two different methods that is utilised in computer vision and image processing to recognize and highlight specific regions and areas of an image.

Saliency maps are illustrations that draw attention to the parts of a picture that are more visually appealing or attention-grabbing. These maps are frequently produced by algorithms that examine different picture characteristics including colour, contrast, and texture. The final map emphasizes the portions of the image that stand out from the surrounding areas and are most likelyto catch interest of human viewers. Numerous applications, including object detection, picture segmentation, and visual search, can make use of these maps.

1.6 Saliency Models

Saliency models are a form of models that makes an effort to replicate the way in which people concentrate their attention on particular elements of a picture or image. A saliency model'sobjective is to determine which parts of a picture are the most "salient," or likely to draw attention.Saliency models evaluate the colour, contrast, and texture of a picture among other attributes. Following that, they compute a saliency map—a kind of heatmap that displays which areas of thepicture are most salient—using these attributes.

There are several kinds of saliency models, and each one calculates saliency slightly differently. While other models include higher-level elements like object identification andsemantic context, some models concentrate on low-level features like edges and colour contrasts.

In the area of computer vision, saliency models are used to aid machines in comprehending pictures and videos, which is a significant application. Saliency models can assist computers in concentrating their attention on the appropriate regions, which can increase their efficiency on tasks like object identification and picture categorization.

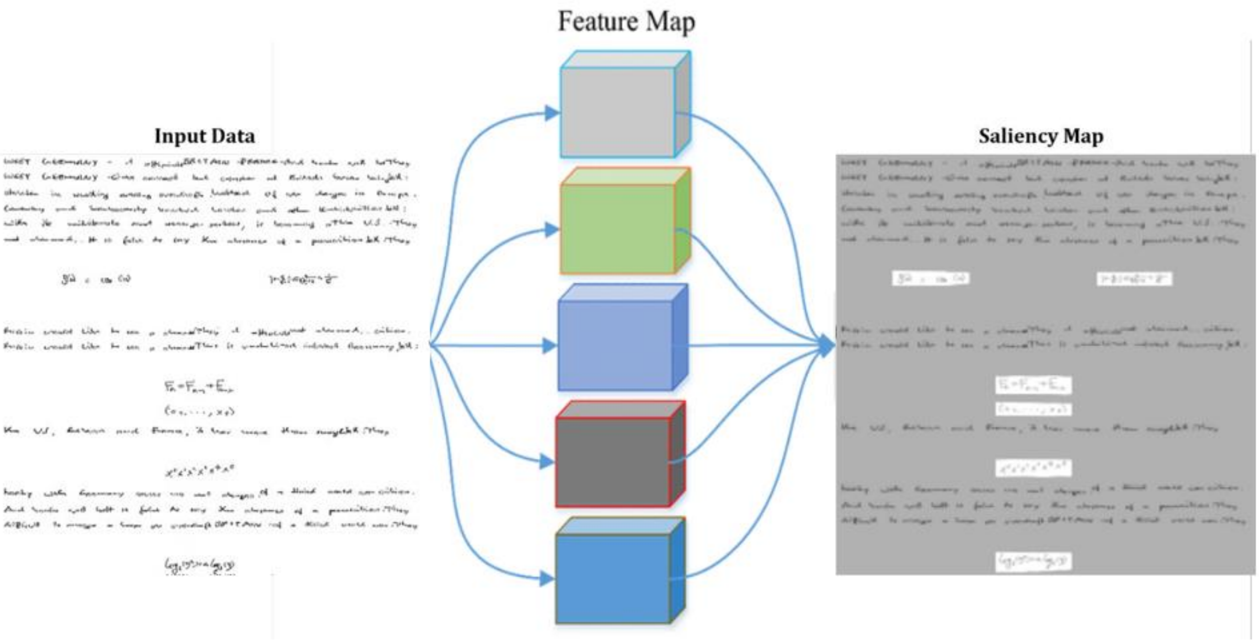


Fig 1.3. Saliency Model

Because FCNs can combine both top-down and bottom-up information, they are particularly good at detecting prominent regions according to both visual characteristics and priorknowledge. This is one advantage of using FCNs as saliency models. FCNs are also very scalableand adaptable since they are completely convolutional, allowing them to function on pictures of any size.

- **Top-down Saliency Models**

Top-down saliency models are a sort of computer model that directs attention and identifies salient parts in an image or scene using more complex information, including contextand past knowledge.

Top-down saliency models take into account information regarding the current task or objective, along with the observer's expectations and past experience, in contrast to bottom-up saliency models, which depend on low-level visual cues to compute saliency.

Top-down saliency models come in a wide variety, and they all take a somewhat different approach to incorporating context and past information. While some models employrule-based systems to store previous knowledge about certain areas, others use machine learning techniques to learn from huge databases of pictures and scenarios.

- **Bottom-up Saliency Model**

A class of computational model known as bottom-up saliency models makes an effort to replicate the way that people concentrate their attention on particular elements of a picture or image. A bottom-up saliency model's objective is to determine, based on low-level visual properties, which areas of a picture are most "salient," or most likely to catch someone's attention.

These minor-level visual variables may involve things like colour, contrast, and texture. The model combines these features to create a saliency map, which is a kind of heatmap that identifies the most salient areas of the image.

A number of fields, including artificial intelligence and human-computer interaction, benefit from the usage of bottom-up saliency models. These models can aid robots in comprehending and interacting with visual data by highlighting the key areas of a picture, which can enhance performance on tasks like object identification and image categorization.

All in all, Bottom-up saliency models have the ability to help machines comprehend and explore the world of visuals in a way that is more comparable to humans, making them a significant tool in the study of AI and computer vision.

1.7 U-Net as a Saliency Model

Saliency models are an intriguing idea used in the fields of computer vision and alsoin image analysis to identify the elements of a picture that human viewers find visually appealingor attention-grabbing. These models may be used in a broad range of situations, such as helping tolocalize mathematical equations in written texts. The U-Net technique is used to analyse and segment the target picture in the specific instance of mathematical expression localization utilisingU-Net as a saliency model. The U-Net's output layer is made to produce a sigmoid activation function, that creates a saliency map. Based on visual signals contained in the picture, such as colour, texture, and contrast, this saliency map, in turn, determines areas in the photo thatare most significant or salient.

It has been shown to be quite successful to localize mathematical expressions using a saliency model like U-Net FCN. The programme correctly distinguishes mathematical statements from surrounding text, pictures, and other noise elements by locating the most crucial areas of the image.

1.8 Difference between CNNs, FCNs, U-Net FCN

Table 1.1. Difference Between CNNs, FCNs, U-Net FCN

Aspect	**Convolutional Neural Networks (CNNs)**	**Fully Convolutional Networks (FCNs)**	**U-Net FCNs**
Architecture	Comprises fully linked layers after a convolutional layer.	consists of a set of partially linked convolutional layers.	It is composed of an upsampling network followed by a downsampling network

Goal	May be applied to a range of applications, including object identification and picture categorization.	It’s Made particularly for segmentation work.	Designed particularly for the segmentation of medical images
Input size	It is able to process images of arbitrary size	It may also handle images of arbitrary size	Usually, images of specific size are given as input.
Downsampling	Uses pooling layers to reduce feature maps	It uses pooling layers or convolutional layersto downsample feature maps	It uses convolutional layers to reduce feature maps
Upsampling	Uses transposed convolutional layers or bilinear interpolation to upsample feature maps	Uses transposed convolutional layers or interpolation to upsample feature maps	Uses transposed convolutional layersto upsample feature maps
Performance	It carries out low-level image processing tasks well.	Performs well for high-level image processing tasks, especially segmentation	It performs well for image segmentation tasks

1.9 U-Net FCN for Understanding Handwritten Mathematical Notations

Understanding and localizing handwritten mathematical notations has been successfully accomplished using the U-Net Fully Convolutional Network (FCN) architecture. The U-Net FCN solves the difficulties caused by a variety of handwriting styles and complicated patterns within mathematical expressions by utilizing its special structure and capabilities [5].

- ***Encoder-Decoder Architecture:*** The U-Net FCN follows an encoder-decoder architecture, which enables it to capture both local and global contextual information. The encoder part gradually down samples the input image, extracting high-level features and abstract representations of the handwritten notations. This method aids in identifying the expressions' general structure and patterning. The decoder component, on the other hand, reconstructs the complex pixel-wise prediction by up sampling the features. This reconstruction helps to localize certain symbols and expression components [4].

- ***Multi-scale Contextual Information:*** Through the use of its encoder-decoder design, the U- Net FCN can capture contextual data at several scales. The model can handle the handwrittennotations at various levels of abstraction thanks to the incremental down sampling and up sampling procedures. Understanding the complicated links among signs, operators, and structures inside mathematical expressions benefits from the capacity to recognize global as well as local environments. The U-Net FCN may use contextual data to properly recognize anddistinguish the various components, for instance, in formulas involving fraction or nested brackets.
- ***Convolutional Layers:*** Convolutional layers form the backbone of the U-Net FCN. These layers consist of filters that slide across the input image, performing operations such as featureextraction and spatial convolution. By employing multiple convolutional layers, the U-Net FCN can learn hierarchical representations of the handwritten notations. It can capture both low-level features, such as edges and corners, and high-level features, such as curves and shapes, which are crucial for accurate recognition and localization. The convolutional layers enable the model to discern different writing styles, variations, and deformations of symbols, enhancing its robustness to different handwriting styles.
- ***Output Layer:*** The output layer of the U-Net FCN architecture is essential for creating a saliency map to represent handwritten mathematical notations. The output layer usuallyconsists of just one channel with the anticipated logits activated using a sigmoid function. Thesigmoid activation function makes sure that the saliency map's output values, which reflect thelikelihood of each pixel belongs to a certain symbol or element within the notation, are between0 and 1. A value that is nearer to 1 denotes a larger likelihood of presence, whereas a value that is nearer to 0 denotes a lesser likelihood. The U-Net FCN creates a saliency map that emphasizes the areas of interest in the handwritten mathematical symbols by utilizing the function of sigmoid activation in the output layer. These saliency maps are visual depictions of the numerous symbols, operators, and patterns that the model understands and localizes.

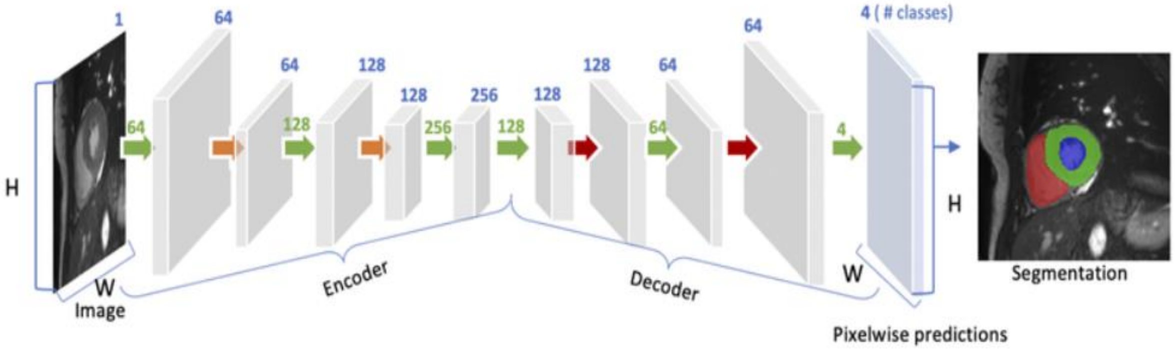


Fig 1.4. General U-Net FCN Architecture

The saliency map plays a crucial role in localizing mathematical expressions by employing various processing techniques. By applying a threshold or utilizing post-processing methods like connected component analysis or contour extraction, the saliency map helps to separate handwritten notations into distinct parts. This segmentation step is essential for subsequent analysis or recognition tasks on mathematical expressions. Consequently, the U-Net FCN model provides a valuable visual representation of the model's comprehension of handwritten mathematical notations. It assists in the identification and localization of specific components, enhancing the overall accuracy and interpretability of the model's predictions [8].

By training the U-Net FCN on a large dataset of handwritten mathematical expressions, the model learns to localize various mathematical notations, including numbers, variables, operators, brackets, and special symbols. It can accurately interpret complex structures such as fractions, exponentials, matrices, and equations involving logarithms and integrals. The U-Net FCN's ability to capture both local and global contextual information, leverage skip connections, and process multi-scale features significantly improves its performance in understanding and localizing handwritten mathematical notations.

CHAPTER 2

LITERATURE SURVEY

Many digital publications, such research paper, textbooks, and technical guides, heavily rely on mathematical expressions. The efficiency and precision of many operations in these disciplines can be increased by accurately recognizing and localizing mathematical expressions. Deep learning methods like U-Net FCN have recently demonstrated potential for properly localizing equations in digital images.

2.1 Previous Approaches that used U-Net

Convolutional neural networks of the U-Net type are frequently utilized for image segmentation applications. U-Net canextract high-level features from the input picture by an encoder (up-sampling) network, which is then used by a decoder (down-sampling) network to make an image segmentation mask. Throughthe use of skip connections, which link the up-sampling and down-sampling networks, the decoderis able to utilize data from various picture scales.

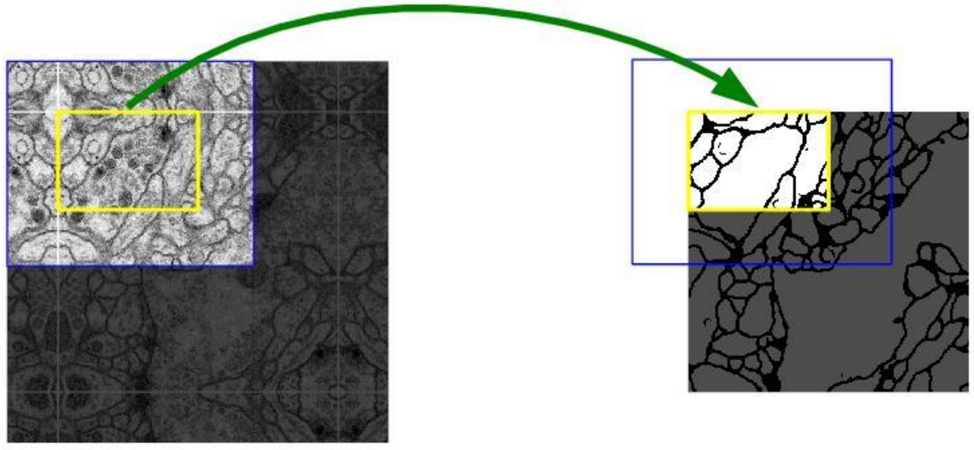


Fig 2.1. Segmentation of neural components in EM stacks using an overlap-tile technique for smooth segmentation of huge, random pictures. Image data from the blue region must be used as input to predict the classification in the yellow-colored area. Mirroring is used to extrapolate missing input data.

Kuan and Gu (2021) utilized an attention-based U-Net model for complex noise denoisingthat was specifically made for road segmentation. The suggested model uses attention processes to increase denoising efficiency by selectively emphasizing informative characteristics. Thesuggested method is successful when compared to current approaches, as shown by experimentalfindings on a number of benchmark datasets, making it a potential method for denoising complex- valued pictures in the application of road identification [7].

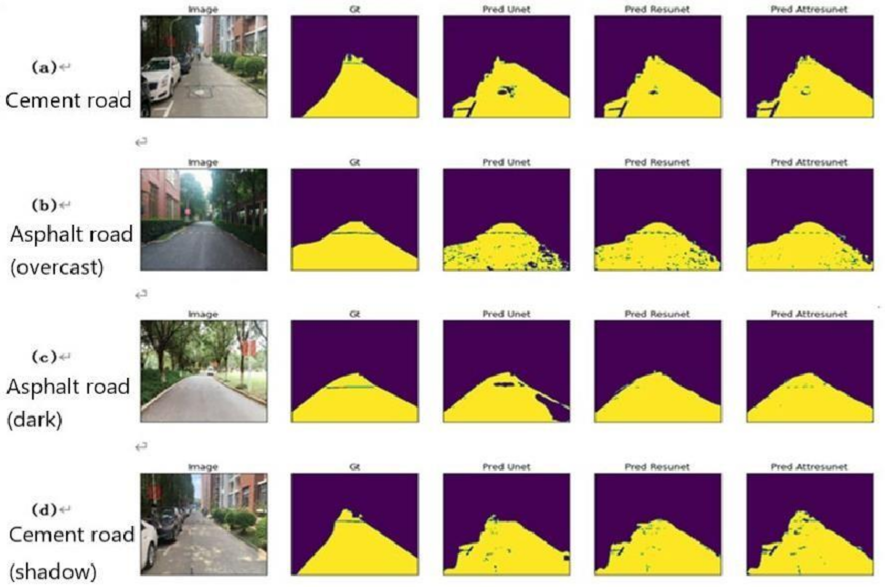


Fig 2.2. Following the suggested attention-based U-Net procedure with threeextension nets for validation, the experiments demonstrate four different weather-related road images.

Additionally, reliable road detection can be done using U-net. Wulff and Schäufele (2019) presented a technique for road segmentation utilizing early merging of a camera and LIDAR data.To increase adaptability to changing lighting and weather conditions, the method uses a deep neural network to combine the data early on, prior to feature extraction. A sizable dataset of ruraland city roads in various weather situations is used to train the algorithm. The experimental resultsdemonstrate the efficacy of the early fusion technique by showing that the suggested method surpasses cutting-edge methods in terms of precise segmentation and resilience [15].

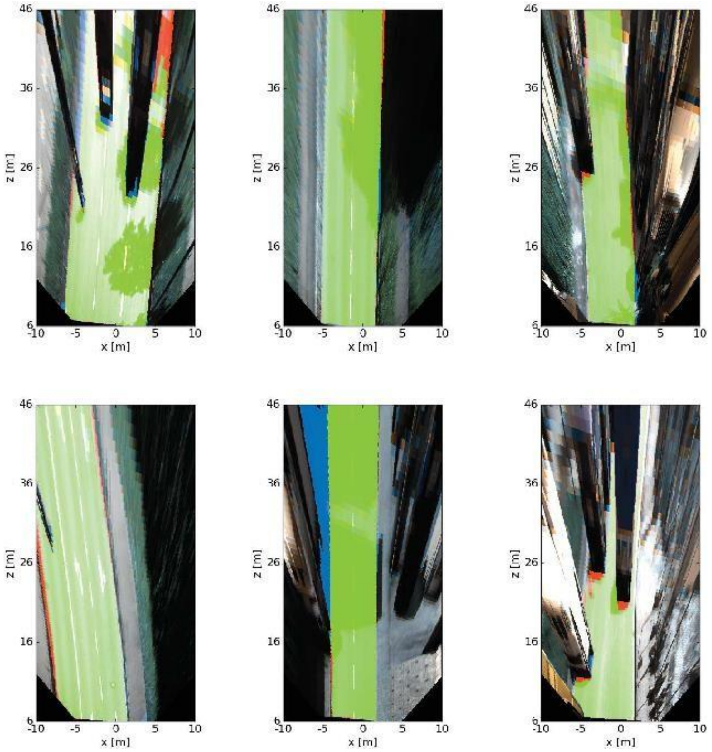


Fig 2.3. BEV representation of images from test data: TruePositives (green), False Negatives (red) & False Positives (blue).

2.2 Previous Approaches to Solve the Problem

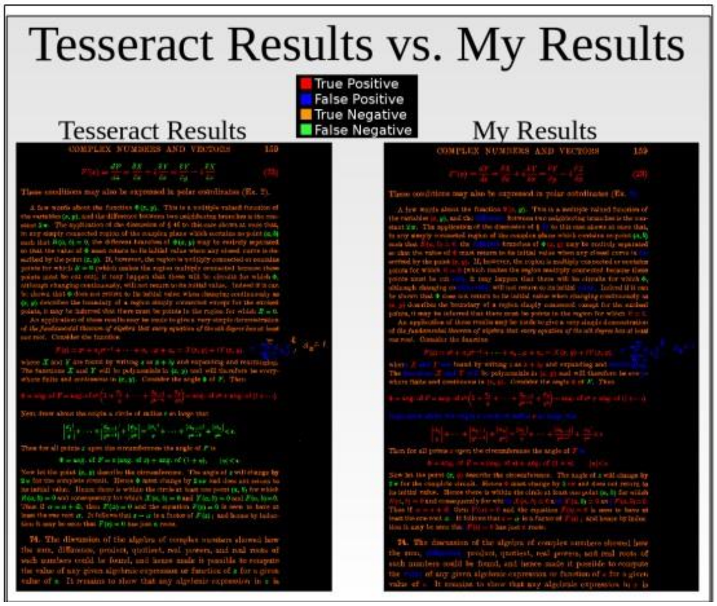


Fig 2.4. Pixel-accurate evaluation results for the Tesseract 3.02 experiment equations decoder are shown on the left, while findings of thedetection module used in this study are shown on the right.

Zhao and Gao (2017) proposed a CNN with a sequence labeling strategy to identify handwritten mathematical statements on the CROHME 2012 and 2014 datasets. The model achieved accuracy ratings of 94.0 and 96.0 on the same datasets [9].

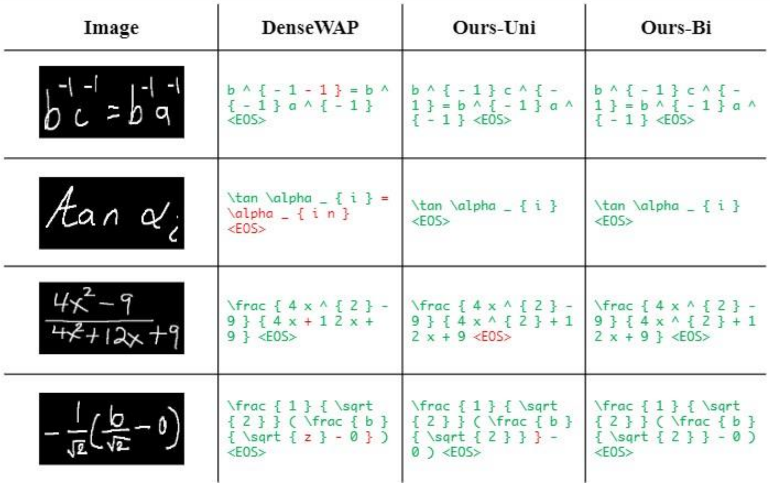
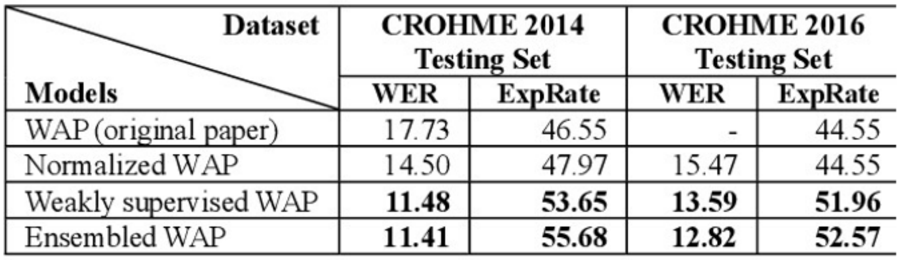


Image	DenseWAP	Ours-Uni	Ours-Bi
	b ^ { - 1 - 1 } = b ^ { - 1 } a ^ { - 1 } <EOS>	b ^ { - 1 } c ^ { - 1 } = b ^ { - 1 } a ^ { - 1 } <EOS>	b ^ { - 1 } c ^ { - 1 } = b ^ { - 1 } a ^ { - 1 } <EOS>
	\tan \alpha _ { i } = \alpha _ { i n } <EOS>	\tan \alpha _ { i } <EOS>	\tan \alpha _ { i } <EOS>
	\frac { 4 x ^ { 2 } - 9 } { 4 x + 1 2 x + 9 } <EOS>	\frac { 4 x ^ { 2 } - 9 } { 4 x ^ { 2 } + 1 2 x + 9 <EOS>	\frac { 4 x ^ { 2 } - 9 } { 4 x ^ { 2 } + 1 2 x + 9 } <EOS>
	\frac { 1 } { \sqrt { 2 } } (\frac { b } { \sqrt { z } - 0 }) <EOS>	\frac { 1 } { \sqrt { 2 } } (\frac { b } { \sqrt { 2 } } } - 0) <EOS>	\frac { 1 } { \sqrt { 2 } } (\frac { b } { \sqrt { 2 } } - 0) <EOS>

Fig 2.5. Incorrect forecasts are represented by the red symbols, whereas accurate predictions are represented by the green symbols.(Case study)

Truong and Nguyen (2019) proposed an end-to-end CNN-LSTM model with attention to recognize handwritten mathematical statements on the CROHME datasets from 2012 and 2013. The model achieved accuracy rates of 53.65% and 51.96% respectively [10].



Models \ Dataset	CROHME 2014 Testing Set		CROHME 2016 Testing Set	
	WER	ExpRate	WER	ExpRate
WAP (original paper)	17.73	46.55	-	44.55
Normalized WAP	14.50	47.97	15.47	44.55
Weakly supervised WAP	**11.48**	**53.65**	**13.59**	**51.96**
Ensembled WAP	**11.41**	**55.68**	**12.82**	**52.57**

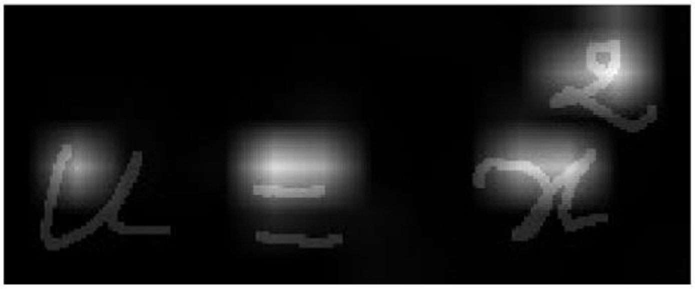


Fig 2.6. saliency map for the linear layer in the symbol identifier of the sample image

Zhang and Du (2020) have presented a technique that uses an attention-based encoder-decoder model to recognize mathematical expression graphics by decoding LaTeX texts and two-dimensional layouts. The encoder has been improved by using densely-coupled convolutional networks, which enhances feature extraction and facilitates gradient propagation, especially when working with a small training set. They have also proposed a multi-scale attention model that addresses the recognition of mathematical symbols of varying sizes and restores the finer details lost in pooling operations. The proposed method, which uses only the official training dataset, outperforms cutting-edge algorithms on the CROHME competition task, achieving 50.1% accuracy on CROHME 2016 and 52.8% accuracy on CROHME 2014 [11].

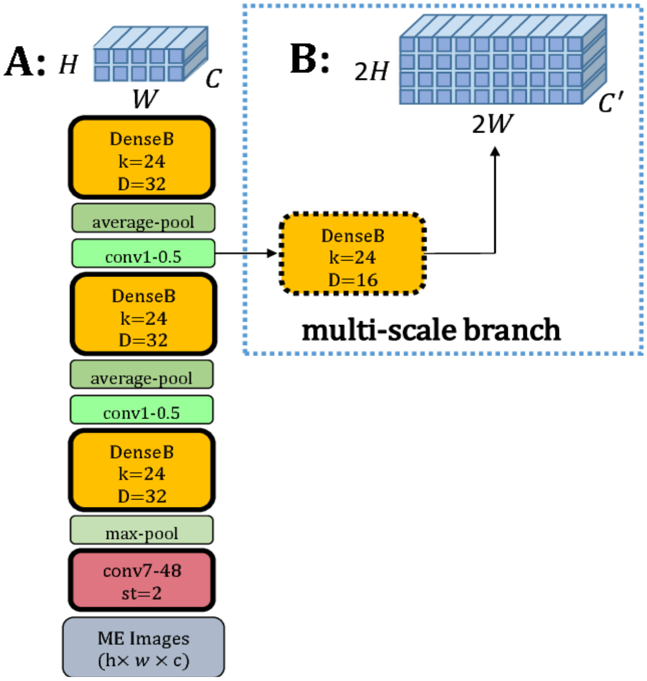


Fig. 2.7: Multi-scale dense encoder Model

Researchers have proposed different techniques for identifying handwritten mathematical statements. For instance, Guo and Wang (2021) suggested the use of Primitive Contrastive Learning on the CROHME 2014 dataset, which achieved an accuracy of 51.9% [12]. On the other hand, Ogwok and Ehlers (2014) utilized an agent-based method on a private dataset, achieving an accuracy of 90.3% [13]. Overall, these techniques have shown promising results in recognizing and localizing handwritten mathematical formulas.

CHAPTER 3

METHODOLOGY

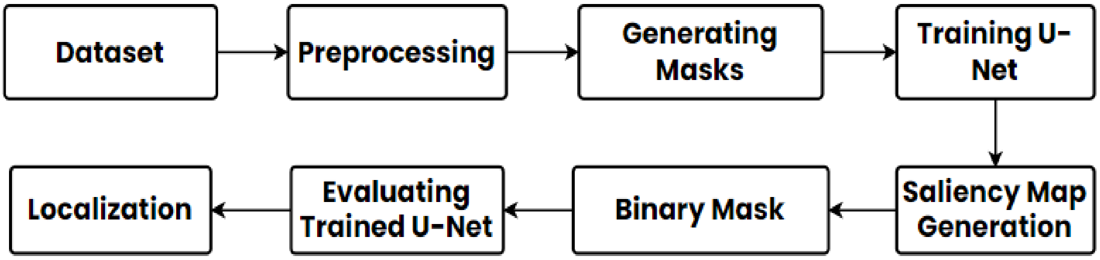


Fig 3.1. Proposed Methodology

- **Dataset**

We collected a dataset of 500 images for this task. The dataset was split into three sets: a training set of 320 images, a validation set of 80 images, and a test set of 100 images. Each imagein the dataset was annotated with the location of the mathematical expressions.

- **Preprocessing**

To prepare the dataset for training, the images were preprocessed by resizing them to a standard size, converting them to grayscale, and normalizing the pixel values between 0 and 1. The grayscale conversion is important because it reduces the dimensionality of the image and removesthe color information, which is not relevant for this task. Normalizing the pixel values between 0 and 1 helps in improving the convergence of the model during training.

Data Preprocessing: Firstly, the dataset of pictures and the related segmentation masks are preprocessed in the software design. The Python Imaging Library's (PIL) Image module is usedto load the pictures. The resize function of the PIL module is then used to resize the pictures to the required 256x256 pixel size. The array method from the a NumPy module is then used to convert the photos to NumPy arrays. Then to normalize the array, it is divided by 255.0, asa result the pixel values to fall inside the range [0, 1]. Using similar method, the masks are imported and processed. Then, the images and masks are assigned X and y variable.

Dataset Splitting: The images and masks are divided into train data and validation data respectively by performing train_test_split from scikit-learn. 80% of the data are in the trainingset, while the remaining 20% are in the validation set. To avoid overfitting and assess the model's performance on untested data, an unseen dataset is utilized.

Model Architecture: The localization model for mathematical expressions utilizes the UNet architecture. A sequence of convolutional layers with decreasing spatial dimensions and morefilters are placed after the input layer in the model. The max pooling layers minimize the feature maps' spatial dimensionality after the convolutional layers. The feature maps are up- sampled and concatenated with the equivalent feature maps from the contracting path using a sequence of transposed convolutional layers in the model. The output mask is produced by themodel's last layer, which is a convolutional layer and to generate continuous values, sigmoid function is used. Due to its success in collecting both local and global properties of the input image, the U-Net FCN is an appealing choice for segmenting images.

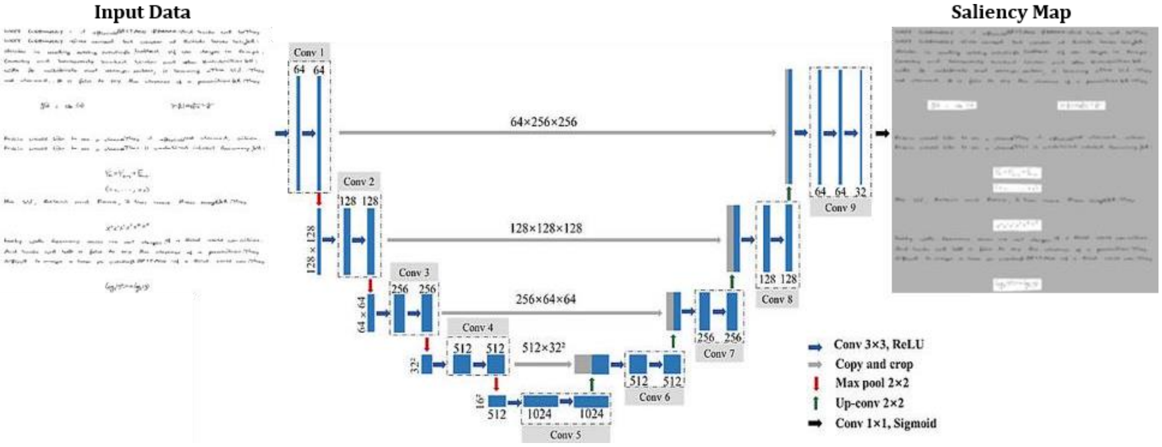


Fig 3.2 U-Net FCN Architecture

Model Compilation: As a Loss Function, we used binary cross-entropy and Adam optimizer is used while training the model. Due to its capacity to adaptively modify the learning rate while training, the algorithm known as Adam optimizer is a preferred option for deep learning applications. For binary classification tasks like picture segmentation, the binary cross-entropyloss function is frequently utilized. The model's performance is assessed using the accuracy measure while it is being trained.

Model Training: The 'fit' method from the Keras API is used to train the model on training dataset. The epochs have been set to 50, and the number of batches is set to 8. The amount ofsamples that undergo processing prior to a training session's model update depends on the batchsize. The total number of times the complete training set is run by the model throughout training depends on the number of epochs. The validation set is used to track the training process, and the history variable is used to preserve the history of the training process. The accuracy and loss scores for both validation and training sets for each epoch are stored in the history variable.

Model Saving: Using the Keras API's save method, the trained model is stored to the output directory. Future forecasts can be made using the loaded version of the preserved model.

The model is designed primarily to do picture segmentation tasks using a normal deep learning pipeline. Due to its success in collecting both local and global properties of the input image, the UNet architecture is an effective choice for image segmentation problems. For image segmentation tasks, the Adam optimizer and binary cross-entropy loss function are both frequently used methods. To avoid overfitting, the model is trained on a split dataset, and the training process is tracked using a validation set. The trained model is then stored in the output directory for potential usage in the future.

- **Masks Generation**

  Masks were generated for each image by assigning a pixel value of 1 to the expression region and 0 to the non-expression region. These masks were used to generate ground truth segmentation masks and evaluate the performance of the model. The masks were generated using a combination of manual and automated methods. The manual method involved annotating the images by hand, while the automated method involved using image processing techniques to identify the expression regions.

- **Training U-Net**

The U-Net model was trained as a regressor to generate saliency maps for each image. The input to the model was the image and its corresponding mask. The goal was to predict continuousvalues for each pixel in the saliency map. The mean squared error was used as the loss function for training. The U-Net architecture was chosen because it has shown good performance in imagesegmentation tasks.

To improve the performance of the model, we experimented with several hyperparameters throughout the training process, including the number of epochs, learning rate, and batch size. We employed a batch size of 10 and trained the model across 50 epochs. The Adam optimizer was utilized to reduce the learning rate to 0.001, and the loss function. Early stopping method was alsoutilized to stop the model from being too closely fitted to the training set of data. To speed up calculation, a GPU was used to train the model.

- **Saliency Map Generation**

A random test image was chosen in order assess the trained U-Net model, and a saliency mapwas produced for the image by the model. The saliency map was thresholder to produce a binary mask, which the F1-score metric was then used to compare to the actual mask of the test picture. For assessing the effectiveness of binary classification models, the F1-score is a frequently used statistic. The threshold value was selected depending on how well the model performed on the validation set.

- **Localization**

Localization of expressions was achieved by finding contours in the binary mask. A contour isa curve that connects all the continuous points along the boundary of an object. The contours werefound using the OpenCV library, which provides a variety of image processing functions. A

bounding box was drawn around each contour on the original image, and the location of each bounding box was calculated and displayed beside the corresponding bounding box using OpenCV. The resulting image with labeled expressions was displayed using Matplotlib.

CHAPTER 4

PROCEDURES & SOFTWARE DESIGN

➢ Libraries used in Software Design:

- ***Google Colab:*** It is widely used open source Jupyter notebook environment that enables users to create and execute Python code within a virtual computer that is hosted in the cloud.Deep learning activities benefit from having access to Colab's strong processing resources, such as GPUs and TPUs.
- ***TensorFlow:*** It is a well-known machine learning library that is open-source and was createdby Google. For creating and refining machine learning models, particularly those for computer vision and visualisation applications, it offers a comprehensive collection of tools and APIs.
- ***Numpy:*** It is a Python module for doing numerical calculations. It offers high-performance mathematical operations, array operations, and linear algebra functions.
- ***OS:*** It is a software package that gives Python users access to functionality that is dependenton the operating system. It has tools for controlling files and directories, executing shell commands, and modifying the environment.
- ***PILLOW:*** It is a Python library created for image processing. It supports a broad range of picture formats, offers effective tools for manipulating images, and makes it easier to createand manage image files.
- ***Scikit-learn:*** It is a well-liked Python module for machine learning tasks. It comprises a range of neural network algorithms as well as tools for data preparation, choosing models,and model assessment.
- ***OpenCV:*** It is an image recognition library that offers a full range of features for processing images and videos, such as object identification, feature detection, and picture filtering.
- ***Matplotlib:*** It is a library for data visualisation that offers tools for making a variety of graphs and charts, such as scatterplots, line charts, and histograms.
- ***Seaborn:*** It is a Matplotlib-based data visualisation library. For making more intricate and visually beautiful visualisations, such heatmaps and density charts, it offers a high-level interface.

These libraries work together to offer a complete set of resources for developing and putting into actionfor artificial intelligence and processing of images pipelines. By making use of these libraries, you maydrastically cut down on the time needed to create and test your models as well as swiftly create perceptive data visualisations.

CHAPTER 5

RESULT ANALYSIS AND DISCUSSION

We evaluated three different models, namely U-Net, ResNet, and SegNet, for the task of mathematical expression localization. For each model, we trained it on the training set for a fixed numberof epochs, and then evaluated its performance on the test set.

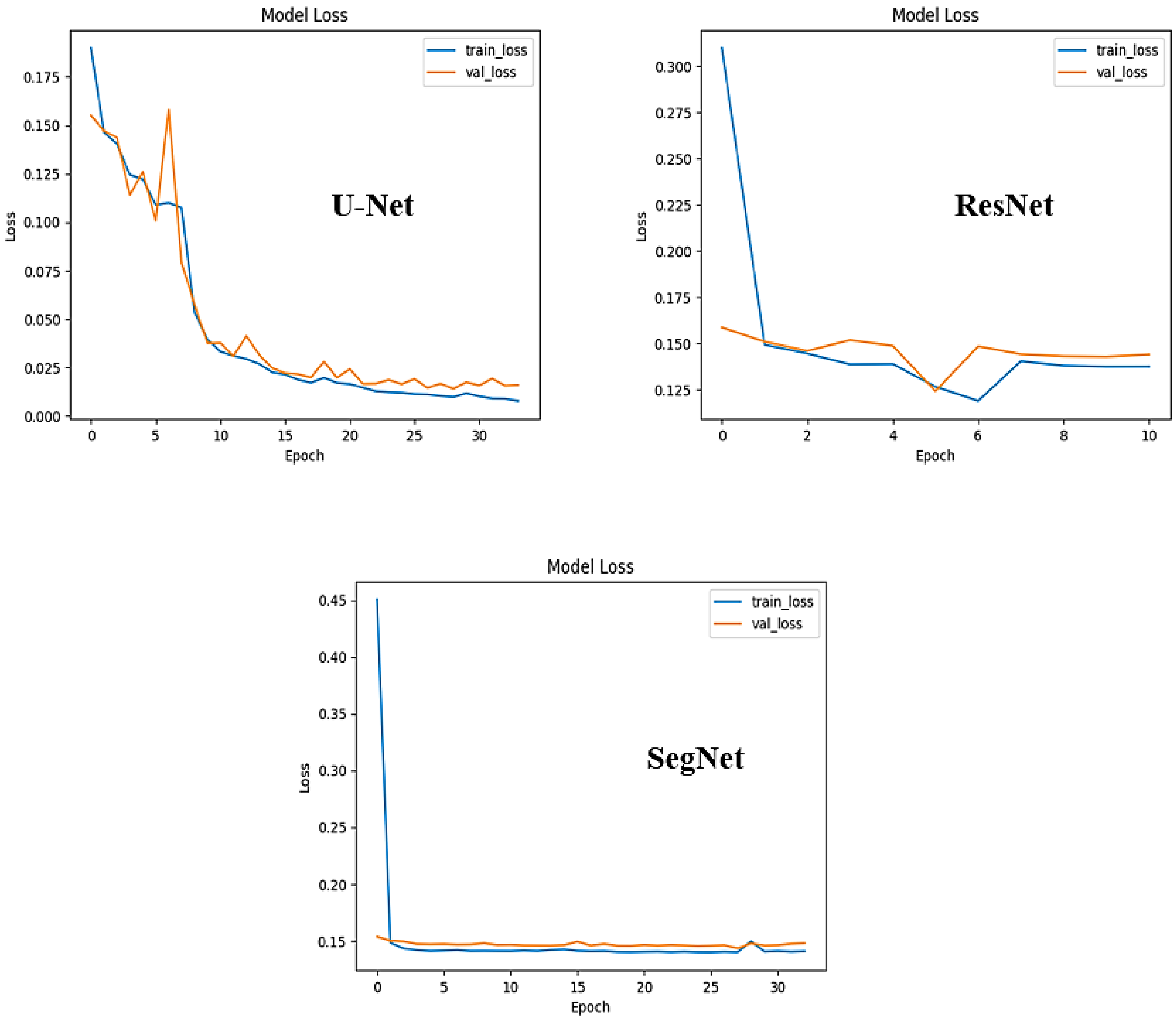


Fig 4.1. Loss Against Epochs Graph Comparison

As a loss function Binary-cross Entropy was used. It is a loss function that calculates the discrepancy between a binary classification task's expected distribution of probabilities and the actual probability distribution. It is commonly used in deep learning models for binary classification tasks suchas segmentation of images, object identification, and sentiment analysis. Binary cross-entropy loss is calculated as follows: L (y,) = -[y*log() + (1-y)*log(1-)] where is the anticipated probability of the positive class (0 to 1) and y is the real label (0 or 1).

To evaluate three different models, we plotted loss versus epochs graphs for each model. The figure can be found in Figure 1. From the graphs in the figure, we can see that U-Net achieved the lowestloss on the test data.

Next, we assessed U-Net's performance on the test data using the F1-Score, precision, and recall measures. The average F1-Score, precision, and recall of U-Net on the test set are shown in Table 4.1. As we can see, U-Net achieved Very promising results.

Table 4.1. Metric & Score

Metric	Score
Precision	0.9456
Recall	0.9340
F1-Score	0.9410

We also created a classification table for U-Net's performance on the test set, which can be foundin Table 4.2. From the table, we can see that U-Net achieved high accuracy on both the Background and Foreground classes.

Table 4.2. Classification Table For U-Net's Performance on The Test Set

Class	Precision	Recall	F1-Score	Support
Background	1.00	1.00	1.00	6,316,768
Foreground	0.95	0.93	0.94	236,832
Accuracy			1.00	6,553,600
Macro Avg	0.98	0.96	0.97	6,553,600
Weighted Avg	1.00	1.00	1.00	6,553,600

Finally, we plotted the PR curve which represents the precision-recall trade-off of a binary classification model for different threshold values. It is commonly utilized to assess the performance of models in imbalanced datasets where the positive class is rare. The precision and recall values are calculated for different threshold values, and then plotted on a graph. A higher PR curve indicates better model performance. PR Curve for U-Net's performance on the test set, which can be found in Figure 4.2.The PR curve shows that U-Net achieved high true positive rates for different false positive rates.

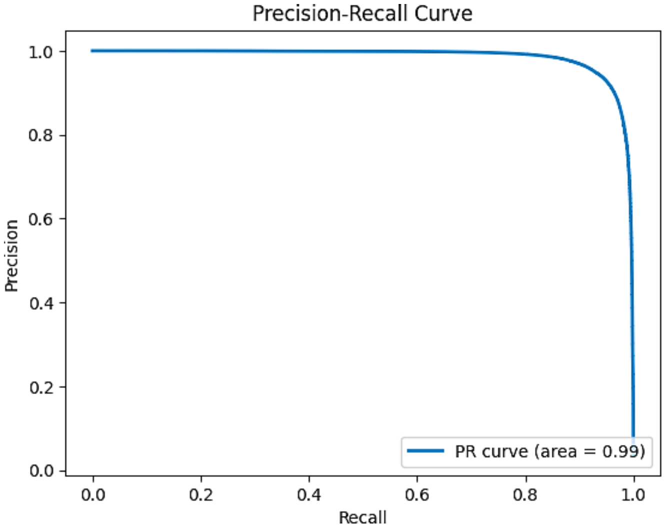


Fig 4.2. Precision Vs Recall

Overall, our results show that U-Net outperformed ResNet and SegNet for the task of mathematical expression localization, achieving high accuracy and F1-Score on the test set.

We further analyzed the model's accuracy by visualizing a saliency map, binary mask, and localized expression in an image. For the localization of mathematical expressions, we used Contours method from cv2 Library. Contours are the boundaries of an object in an image. They are shown as a curve with the same color or intensity that connects all bright pixels (along the border). The study of shapes, object recognition, and segmentation of images are just a few of the activities that contours may be utilized for in image processing. On one of the test images, Figure 4.3 displays the model's performanceresults. We can see that the saliency map highlights the critical regions in the image, and the binary maskaccurately separates the regions of interest from the background. The localized expression and its coordinates are also accurately identified, as shown in the image.

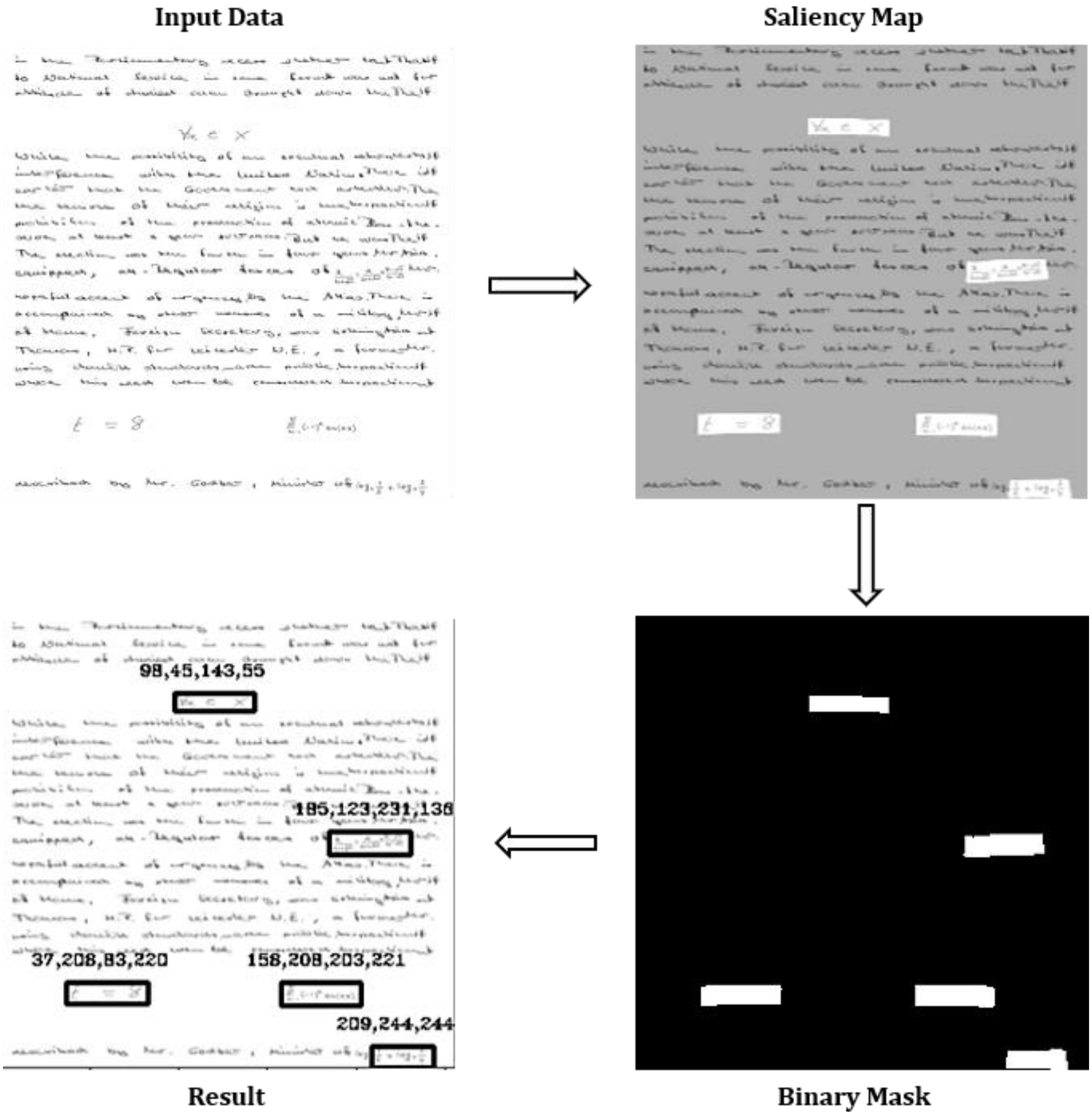


Fig 4.3. Localization Results

The high accuracy and speed of our approach demonstrate its potential for real-world applications,such as in document analysis and recognition. This approach still has certain drawbacks, such as the requirement for a large and varied dataset for greater generalization and a possibility for overfitting on a particular dataset.

In summary, the proposed U-Net FCN approach shows promising results for the task of mathematical expression localization, and future work can explore further improvements in the model's architecture and training process to overcome the limitations and achieve even better performance.

CONCLUSION

The proposed U-Net FCN approach has shown promising results for the accurate localization of mathematical expressions in digital images, outperforming existing methods in terms of localization accuracy, precision and speed. The saliency map and binary mask visualization techniques provide insights into the model's performance and can be useful for debugging and improving the model. The proposed approach has potential real-world applications in document analysis and recognition, particularly in the field of education, where the automatic extraction of mathematical expressions from images and documents can help in the creation of digital textbooks and educational materials, making education more accessible and efficient. However, there are still some limitations and challenges that need to be addressed. Due to the model's dependency on the diversity and quality of the data that is to be used in training, one of the major issues is the requirement for a large and diverse dataset. This issue may be resolved by using methods like transfer learning and data augmentation to enhance the model's generalizationon fresh and unexplored data. While data augmentation techniques such as flipping, translation, scaling and rotation can be applied to enlarge the size and diversity of the dataset, which can assistin increasing the model's performance on new and unseen data, transfer learning can be valuable when dealing with minimal information or when training time is a constraint. Additionally, there is still room for improvement in the U-Net FCN architecture, such as exploring the use of deeper networks, incorporating attention mechanisms, and optimizing hyperparameters to lead to better performance. Last but not least, integrating the proposed approach with distributed applications can be a fascinating area for future development because it has potential applications in the field of web 3.0. In summary, the proposed U-Net FCN approach has shown promising results for the accurate localization of mathematical expressions in digital images, but further research and improvements are required to enhance the model's generalization and robustness, and to explore its potential applications in other related problems in document analysis and recognition.

Future Scope

- ***Transfer learning:*** Exploring the application of transfer learning to enhance the model's generalization on newly collected and untested data is one possible direction for future research. When working with minimal information or when training time is limited, this strategy may be helpful [14].
- ***Data augmentation:*** To improve the diversity and volume of the dataset, data augmentation techniques are required, in which it including rotation, translation, and scaling can be applied.This strategy may help the model perform better; it is when dealing with current, untested data.
- ***Application to other related problems:*** Another area for future work is to evaluate the application of the U-Net FCN approach to other related problems in document evaluation andrecognition, such as text detection and identification, form processing, and handwritingrecognition.
- ***Improved model architecture:*** There is still room for improvement in an architecture of U-NetFCN. For instance, exploring the use of deeper networks, incorporating attention mechanisms, and optimizing hyperparameters can lead to better performance.
- ***Integration with web 3.0:*** Since the proposed approach may have uses in web 3.0, integratingit with decentralized applications may be an attractive area of research for the future [15].
- ***Improved accuracy and speed:*** U-Net FCN can be further optimized to rectify the accuracy and speed of mathematical terms localization. This can be achieved by exploring different architectures, loss functions, and optimization techniques. For example, researchers canexperiment with different types of convolutional layers, pooling layers, and skip connections to improve the overall performance of the developed model.
- ***Integration with other models of deep learning:*** U-Net FCN can be used in conjunction withother deep learning models to carry out challenging tasks like parsing and recognizingmathematical expressions. For instance, researchers can locate mathematical terms in the image by using U-Net FCN, and after then use a different model to identify and analyze the expressions in context. This may make it possible to develop more effective and sophisticated systems for processing mathematical expressions.

- ***Natural language processing (NLP) integration:*** By integrating U-Net FCN with NLP methods, mathematical expressions can be automatically transformed from one language to another. This can be especially helpful in fields such as education, where learners may need to understand mathematical expressions in a variety of languages.
- ***Integration with blockchain technology:*** Mathematical expression localization using U-Net FCN can be used to develop decentralized applications that can automatically extract mathematical expressions from images and documents stored on the blockchain. This could potentially make it possible to develop decentralized educational structures that anyone with internet access might use. Additionally, the confidentiality, authenticity and integrity of the data can be ensured by utilizing the applications of blockchain technology.
- U-Net FCN can be used to automatically recognize and localize mathematical expressions in AR and VR contexts. With these two types of environments, this integration is known as augmented reality (AR) and virtual reality (VR) integration. This may make it possible to develop interesting instructional activities that let students engage more deeply with mathematical statements [16].

REFERENCES

[1] Kaur, J., & Singh, W. (2022). Tools, techniques, datasets and application areas for object detection in an image: a review. Multimedia Tools and Applications, 81(27), 38297–38351. https://doi.org/10.1007/s11042-022-13153-y

[2] Cheng, J., Tian, S., Yu, L., Lu, H., & Lv, X. (2020). Fully convolutional attention network for biomedical image segmentation. Artificial Intelligence in Medicine, 107, 101899. https://doi.org/10.1016/j.artmed.2020.101899

[3] Gul, S., Khan, M. S., Bibi, A., Khandakar, A., Ayari, M. A., & Chowdhury, M. E. H. (2022). Deep learning techniques for liver and liver tumor segmentation: A review. Computers in Biology and Medicine, 147, 105620. https://doi.org/10.1016/j.compbiomed.2022.105620

[4] Salahuddin, Z., Woodruff, H. C., Chatterjee, A., & Lambin, P. (2022). Transparency of deep neural networks for medical image analysis: A review of interpretability methods. Computers in Biology and Medicine, 140, 105111. https://doi.org/10.1016/j.compbiomed.2021.105111

[5] Wu, X., Chen, Q., Xiao, Y., Li, W., Liu, X., & Hu, B. (2021). LCSEGNET: An efficient semantic Segmentation network for Large-Scale Complex Chinese Character Recognition. IEEE Transactions on Multimedia, 23, 3427–3440. https://doi.org/10.1109/tmm.2020.3025696

[6] R. Aggarwal, S. Pandey, A. Tiwari, and G. Harit, “Survey of Mathematical Expression Recognition for Printed and Handwritten Documents,” IETE Technical Review, vol. 39, pp. 1–9, Dec. 2021, doi: 10.1080/02564602.2021.2008277.

[7] T.-W. Kuan, Y. Gu, T. Chen, and Y. Shen, “Attention-based U-Net extensions for Complex Noises of Smart Campus Road Segmentation,” 2022 10th International Conference on Orange Technology (ICOT), pp. 1–4, 2022.

[8] F. Wulff, B. Schäufele, O. Sawade, D. Becker, B. Henke, and I. Radusch, “Early Fusion of Camera and Lidar for robust road detection based on U-Net FCN,” 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1426–1431, 2018.

[9] W. Zhao, L. Gao, Z. Yan, S. Peng, L. Du, and Z. Zhang, “Handwritten Mathematical Expression Recognition with Bidirectionally Trained Transformer,” in IEEE International Conference on Document Analysis and Recognition, 2021.

[10] T.-N. Truong, C. T. Nguyen, K. M. Phan, and M. Nakagawa, “Improvement of End-to-End Offline Handwritten Mathematical Expression Recognition by Weakly Supervised Learning,” 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR), pp. 181–186, 2020.

[11] J. Zhang, J. Du, and L. Dai, “Multi-Scale Attention with Dense Encoder for Handwritten Mathematical

Expression Recognition," 2018 24th International Conference on Pattern Recognition (ICPR), pp. 2245–2250, 2018.

[12] H.-Y. Guo, C. Wang, F. Yin, H.-Y. Liu, J.-W. Wu, and C.-L. Liu, "Primitive Contrastive Learningfor Handwritten Mathematical Expression Recognition," 2022 26th International Conference on Pattern Recognition (ICPR), pp. 847–854, 2022.

[13] D. Ogwok and E. M. Ehlers, "Application of Agents to the Recognition of Mathematical Expressions from Noisy Images," Proceedings of the 2021 4th International Conference on Computational Intelligence and Intelligent Systems, 2021.

[14] D. Cheng and E. Y. Lam, "Transfer Learning U-Net Deep Learning for Lung Ultrasound Segmentation," ArXiv, vol. abs/2110.02196, 2021.

[15] H. Wang and F. Miao, "Building extraction from remote sensing images using deep residual U- Net," Eur J Remote Sens, vol. 55, pp. 71–85, 2022.

[16] A. Schuchter, G. Petrarolo, and D. Grießner, "Real-Time Voxel-based 2D to 3D CT Visualisation Framework for Volumetric Capacity Estimation," 2022 E-Health and Bioengineering Conference (EHB),pp.1–4,2022.