FDRE TECHNICAL AND VOCATIONAL TRAINING INSTITUTE (FTVTI)

DEPARTMENT OF INFORMATION TECHNOLOGY

THESIS TITLE:

AI INTEGRATED ASSISTIVE TECHNOLOGY FRAMEWORK TO ENHANCE LEARNING ACCESSIBILITY FOR VISUALLY IMPAIRED STUDENTS.

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

BY

HAILEMARIAM DERESE MEKURIA
Student ID: TTMR/120/16

Principal Advisor: Getinet Yilma (PhD)

JUNE 9, 2026

ADDIS ABABA, ETHIOPIA

FDRE TECHNICAL AND VOCATIONAL TRAINING INSTITUTE

COLLEGE OF GRADUATE STUDIES

DEPARTMENT OF INFORMATION TECHNOLOGY

AI INTEGRATED ASSISTIVE TECHNOLOGY FRAMEWORK TO ENHANCE LEARNING ACCESSIBILITY FOR VISUALLY IMPAIRED STUDENTS.

By: Hailemariam Derese Mekuria
IDNo: TTMR/120/16

A thesis submitted to the Department of Information Technology, College of Graduate Studies, FDRE Technical and Vocational Training Institute (FDRE TVTI), in partial fulfillment of the requirements for the degree of Master of Science in Information Technology.

Principal Advisor: Getinet Yilma (PhD)

JUNE 9, 2026

ADDIS ABABA, ETHIOPIA

DECLARATION

I, declare that this thesis entitled**:** "AI Integrated Assistive Technology framework to Enhance Learning Accessibility for the Visually Impaired Students" is my original work. I have undertaken the research work independently with the guidance and support of the research advisor. This study has not been submitted for any degree or diploma program in this or any other institutions and that all sources of materials used for the thesis has been duly acknowledged.

Declared by

Name: Hailemariam Derese Mekuria

Signature: _____

Department: Information Technology, FDRE TVTI.

Date: _____

This dissertation has been submitted for examination with my approval as the candidate's advisor.

Principal Advisor: Dr. Getinet Yilma

Signature: 

Date: _____

APPROVAL SHEET

This is to certify that the thesis entitled ***"AI Integrated Assistive Technology framework to Enhance Learning Accessibility for the Visually Impaired Students"*** submitted by ***Hailemariam Derese Mekuria with IDNo: TTMR/120/16,*** in partial fulfillment of the requirements for the degree of Master of Science in Information Technology, has been examined and approved by the board of examiners of the department of Information Technology, FDRE Technical and Vocational Training Institute.

Dr. Getinet Yilma		_____
Name of advisor	Signature	Date
_____	_____	_____
Internal Examiner	Signature	Date
_____	_____	_____
External Examiner	Signature	Date
_____	_____	_____
Department Head, IT	Signature	Date
_____	_____	_____
Dean, College of Graduate studies	Signature	Date

Place: Addis Ababa, Ethiopia

Date of Final Approval: _____

ACKNOWLEDGMENT

First of all, I thank Almighty God and St. Mary, Mother of our Lord, for the strength and wisdom to do this research and for their blessings every step of the way!

Then, I wish to express my deep gratitude to my advisor **Dr. Getinet Yilma** for his support, insight, comments, encouragement, and patience. I am deeply grateful for his advice and guidance from the time I proposed the study to the conclusion of this study. This thesis could not have been accomplished without his valuable feedback and corrections.

Furthermore, my sincere thanks go to all my classmates, who generously sacrificed their study time to answer my questions and provide explanation whenever needed.

Finally, I am truly grateful to my mother, Belaynesh Kebede. She has always supported me with her prayers and support to make sure I do well in my postgraduate studies.

ABSTRACT

Inclusive Education is a foundation to most modern education programs that promotes equal learning opportunity access for all students, regardless of the differences they may have based on physical or cognitive characteristics to a higher quality education. This study explores the ability of assistive technology to read visual instructional content in education. Screen readers such as JAWS can read text, but cannot read diagrams, charts, and visual instructional content, such as diagrams, charts, and graphics, that is typically inaccessible to visually impaired Students. The study proposes JAIF through Design Science Research. It is based on a YOLOv12-based object detection model to recognize diagram elements, Optical Character Recognition (OCR) to extract text, Natural Language Processing (NLP) to generate semantic descriptions, and a Text-to-Speech module in JAWS that produces real-time audio output. The system was tested using 200 STEM diagrams from textbooks and open educational image for training and evaluation. The system was tested for precision, recall, F1, mean Average Precision (mAP@50 and mAP@50–95), Character Error Rate (CER), Word Error Rate (WER), BLEU, and ROUGE. Results indicated a precision of 0.768, recall of 0.735, mAP@50 of 0.815, mAP@50–95 of 0.530, CER of 0.099, WER of 0.180, BLEU score of 0.116, and a ROUGE score of 0.505. The proposed framework can identify, interpret, and describe visual educational content in real time to make it accessible to visually impaired Students. The study recommends expanding the framework to support multilingual content and testing the system in real classrooms to improve the usability of the system.

Keywords: *Artificial Intelligence, Assistive Technology, Inclusive Education, Accessibility, Computer Vision, Natural Language Processing*

LIST OF ACRONYMS

AI	Artificial Intelligence
JAIF	JAWS-AI Integrated Framework
JAWS	Job Access With Speech
OCR	Optical Character Recognition
NLP	Natural Language Processing
TTS	Text-to-Speech
YOLO	You Only Look Once
CNN	Convolutional Neural Network
DSR	Design Science Research
BLEU	Bilingual Evaluation Understudy
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
mAP	Mean Average Precision
IoU	Intersection over Union
CER	Character Error Rate
WER	Word Error Rate
SUS	System Usability Scale
API	Application Programming Interface
GPU	Graphics Processing Unit
CPU	Central Processing Unit
UI	User Interface
UNCRPD	UN Convention on the Rights of Persons with Disabilities
CV	Computer Vision
DL	Deep Learning

TABLE OF CONTENTS

Table of Contents

DECLARATION I
APPROVAL SHEET II
ACKNOWLEDGMENT III
ABSTRACT IV
LIST OF ACRONYMS V
TABLE OF CONTENTS VI
LIST OF FIGURES IX
LIST OF TABLES IX
CHAPTER 1: INTRODUCTION 1
1.1. Background of the Study 1
1.2. Motivation of the Study 4
1.3. Statement of the Problem 5
1.4. Objective of the study 6
1.4.1. General Objective 6
1.4.2. Specific Objectives 6
1.5. Research Questions 7
1.6. Scope and limitation of the Study 7
1.7. Significance of the Study 8
1.8. Organization of the Thesis 9
CHAPTER 2: REVIEW OF RELATED LITERATURE 10
2.1. Introduction 10
2.2. Theoretical Literature Review 11
2.2.1. Overview of Deep Learning and Object Detection 11
2.2.2. Assistive Technology in Inclusive Education 11
2.2.3. Screen Readers and Digital Accessibility (JAWS) 12
2.3. Empirical Literature Review 13
2.3.1. Evolution of YOLO Algorithms in Assistive Contexts 13
2.3.2. AI for STEM-Specific Recognition and Diagramming 14
2.3.3. Integration with Screen Readers and User Experience 15
2.3.4. Local Context and Accessibility Barriers in Ethiopia 15

2.4. Related Works 17
2.5. Research Gap 19
2.6. Conceptual Framework 20
CHAPTER 3: RESEARCH METHODOLOGIES 23
3.1. Introduction 23
3.2. Research Design and Approach 23
3.3. Model Architecture 24
3.4. Model Selection and Justification 26
3.4.1. Object Detection Model (YOLOv12) 26
3.4.2. Optical Character Recognition (OCR) 26
3.4.3. Natural Language Processing (NLP) Model 27
3.4.4. Text-to-Speech (TTS) Model 27
3.4.5. JAWS Integration Layer 27
3.5. Data Collection 27
3.6. Data Collection Instruments 28
3.7. Data Preprocessing and Preparation 29
3.8. Data Split 30
3.9. Model Training and Feature Extraction 31
3.10. Libraries and Software Tools 32
3.11. Performance Evaluation Method 33
CHAPTER 4: SYSTEM IMPLEMENTATION AND RESULTS 36
4.1. Introduction 36
4.2. System Development Overview 36
4.3. Dataset Preparation and Analysis 39
4.4. Model Training and Performance Evaluation 42
4.4.1. Training Behavior 42
4.4.2. Detection Performance Metrics 44
4.4.3. Class-wise Performance Analysis 45
4.4.4. Confusion Matrix Analysis 45
4.4.5. Precision–Recall and Confidence Analysis 48
4.4.6. OCR Performance Metrics 51
4.4.7. NLP Performance Metrics 54
4.5. Qualitative Results and Visual Evaluation 57

4.6. System Integration and Assistive Output 58
4.7. Discussion 61
CHAPTER 5: CONCLUSION AND RECOMMENDATION 63
5.1. Conclusion 63
5.2. Contributions of the Study 64
5.3. Limitations of the Study 64
5.4. Recommendations for Future Work 65
References 67

LIST OF FIGURES

Figure 1.Conceptual Framework..... 22
Figure 2.Proposed Model of Architecture for JAIF prototype..... 25
Figure 3.Data Preprocessing and Preparation..... 30
Figure 4.System Architecture of the Proposed JAIF Assistive Framework..... 37
Figure 5.Dataset Split Distribution..... 39
Figure 6.Distribution of Annotated Objects per Class..... 40
Figure 7.Sample Annotated Images..... 41
Figure 8.Training and Validation Curves..... 43
Figure 9.Normalized Confusion Matrix..... 46
Figure 10.Confusion Matrix..... 47
Figure 11.Precision–Recall Curve showing the relationship between precision and recall across different classes..... 48
Figure 12.Confidence Curve illustrating the optimal confidence threshold for balanced detection performance..... 49
Figure 13.Precision–Confidence Curve demonstrating the increase in precision with higher confidence thresholds..... 50
Figure 14.Recall–Confidence Curve showing the decrease in recall as the confidence threshold increases..... 51
Figure 15.OCR Extraction and Assistive Audio Outputs..... 54
Figure 16.Generated Accessible Descriptions..... 56
Figure 17.Sample qualitative detection results showing model predictions on diverse STEM diagrams, including symbols, text regions, arrows, and objects..... 57
Figure 18.System interface showing image upload and detected diagram components..... 59
Figure 19.OCR extraction, generated accessible description, and audio output from detected diagram elements..... 60

LIST OF TABLES

Table 1 Summary of Reviewed Literature..... 16
Table 2.Comparative Analysis of Related Works..... 20
Table 3.Training Configuration..... 42
Table 4.Training Loss Behavior..... 42
Table 5.Overall Detection Performance..... 44
Table 6.Class-wise Detection Performance..... 45
Table 7.Sample OCR Evaluation Results..... 52
Table 8.OCR evaluation results..... 53
Table 9.Summarized OCR evaluation metrics..... 53
Table 10.Sample NLP Evaluation Results..... 54
Table 11.Summarized NLP evaluation result..... 55

CHAPTER 1: INTRODUCTION

1.1. Background of the Study

Inclusive Education is a basic foundation to most modern education programs that support equal access for all students, regardless of the differences they may have based on physical or cognitive characteristics to a higher quality education[1]. This principle is based on global commitments from documents such as “The Salamanca Statement” (1994) and “The UN Convention on the Rights of Persons with Disabilities” (UNCRPD) that promote inclusive education through education systems that accommodate all learner’s diverse learning needs and facilitate all learner’s full participation[2]. It is important to understand that simply placing an individual into an inclusive classroom does not constitute inclusion as there are many other factors (i.e., Accessible Learning materials) that must be in place that lead to an inclusive learning environment. These include technologies that support a fully accessible Learning Environment.

While these global principles of providing inclusive education exist, it is clear that the implementation of inclusive education has been very limited and is even more so in developing nations where education systems currently lack sufficient technology, infrastructure, and adequate resources to support the implementation of inclusive education[3]. In the majority of contexts (developing countries) inclusive education is restricted to physical integration as compared to providing accessible resources and/or instructions (e.g., teaching methods). Additionally, incorporating adaptive technology, accessible instructional materials/content, and inclusive pedagogical strategies that support active participation would produce a more inclusive educational experience for all learners and would ultimately support the achievement of all learners[4].

Persons with disabilities, particularly those who are visually impaired Students, experience many barriers to accessing STEM-based educational activities[5]. For individuals with moderate to severe visual impairments, the number of individuals worldwide affected is estimated to be around 295 million, including approximately 43 million who are completely blind[6]. Studies have been conducted on the completion of postsecondary education among visually impaired Students individuals and find that only about one out of every three visually impaired individuals finishes a

postsecondary education degree, a situation largely caused by inaccessibility in the learning environment[7].

The visual nature of STEM content is a significant contributor to the disparity between visually impaired individuals and traditionally sighted individuals in educational outcomes. Hard-to-imagine concepts in physics, statistical analysis, and engineering are primarily shown visually via diagrams, charts, circuit diagrams, and other graphical representations that are necessary for someone to understand complicated concepts and abstract ideas[8]. The use of these types of graphs does not provide individuals with accessible ways of perceiving this type of content, which hinders students' ability to understand and participate in STEM-related areas of study.

Many assistive technologies have made some improvement to accessibility. Screen readers (JAWS, NVDA) allow visually impaired Students individuals to read the digital text using their voice output or via a Braille display[9]. While these tools work extremely well for structured text, they are severely limited when it comes to interpreting all kinds of non-structured, non-textual visual information. In fact, studies have shown that while the accuracy of screen readers is nearly 100% with respect to text, the accuracy drops drastically (often between 12%-34%) with respect to diagrams and graphics [10].

To overcome these limitations, many different types of solutions have been proposed and some have been used, including tactile graphics and alternative text created manually. Tactile graphics provide someone to use their hands to explore visual content, while alternative text provides descriptive information about images. These two solutions do have limitations regarding the ability to be reproduced on a larger scale and for the ability to offer interactivity and detail. Tactile graphics require a specific piece of equipment and can be very costly to create[11]. while alternative text descriptions are often simply static representations of images and do not adequately illustrate how complex relationships exist within diagrams. Thus, many visually impaired students are dependent on human assistance when learning and, therefore, lack independence and are not actively engaged in learning [12].

Limited technological infrastructure and inadequate access to cutting-edge assistive devices make the barriers even greater in Ethiopia[13]. While there are inclusive education policies implemented, they do not have sufficient resources and a lack of specialized support systems,

which restricts their effective implementation. Thus, the barriers to STEM education encountered by visually impaired students create limited opportunities for them to participate in this type of education and workforce development[14]. Addressing these barriers is important not only for guaranteeing access to education for all but also for creating national growth and building workforce capacity[15].

The rapid advancement of Artificial Intelligence (AI), particularly in Deep Learning-based Computer Vision and Natural Language Processing (NLP), has created promising opportunities for enhancing accessibility for visually impaired Students [16]. These technologies allow machines to perform analysis on visual data, detect patterns and create descriptive outputs. However, the majority of AI systems developed to date are typically general purpose and do not possess the domain-specific attributes that would enable them to accurately interpret complex educational diagrams[17]. Furthermore, most AI systems do not capture the relationships between concepts that provide the foundation for effective learning in a STEM diagrams[18].

The development of advanced object detection models such as YOLOv12 with high speed and accuracy in real-time visual analysis address these limitations[19]. However, the degree to which educators can use these tools was depend on the ability of these systems to integrate with assistive technology that can provide users with useful outputs. This indicates the importance of creating both a unified and structured solution.

This study proposes the design of the JAWS–AI Integrated Assistive Technology Framework (JAIF), which consists of the following components: Computer Vision for diagram detection and text extraction; and Natural Language Processing for generating context-specific descriptions**.** These components are integrated with screen reader technology to provide users with real-time auditory interpretation of visual educational material[20]. In transforming traditional screen reader into intelligent systems, which are able to interpret and understand complex visual interfaces, demonstrate a significant advancement in assistive technology. Additionally, this framework aligns with global initiatives, including the United Nations' Sustainable Development Goals (SDG4) to promote universal access to inclusive, equitable, high-quality education[21]. Thus, through bridging the visual content and auditory accessibility divide, the JAIF framework

promotes increased accessibility to education, and increased independence within an inclusive education system.

1.2. Motivation of the Study

The ongoing motivation for this research stems from the constant barriers to accessing visually related educational material for visually impaired Students. The fields of Science, Technology, Engineering and Mathematically require a lot of visual information but visually impaired Students cannot access these materials due to the limited ability of existing assistive technologies to interpret complex visual elements such as images, graphics, and illustrations.

While existing assistive tools have significantly improved access to textual content via screen readers such as JAWS, there is still a major gap in the ability of assistive tools to process and convey non-textual, visual content to users in real-time. As such, visually impaired students commonly have incomplete knowledge of important forms of instructional content.

Current alternative approaches to assistive technology such as tactile graphics and printed descriptive text provide partial assistance but are expensive to produce, lack scalability and do not enable users to access related or real-time forms of learning. As a result, these learners are often dependent on outside assistance/able-bodied personnel and do not have the ability to be independent or active participants in learning activities.

Recent breakthroughs in Artificial Intelligence (AI) and Computer Vision, Optical Character Recognition and Natural Language Processing create new ways to overcome limitations associated with the traditional methods of providing accessibility. Most current solutions are very fragmented with no common framework that can facilitate seamless integration of the various forms of assistive technology currently being utilized in educational settings.

The goal of this research study is to develop a comprehensive and integrated solution to help bridge the gap between visual content and auditory accessibility. The proposed JAWS-AI Integrated Assistive Technology Framework (JAIF) is a framework that integrates AI ability and screen-reader technology for providing real-time, organized, and contextually aware auditory descriptions of visual educational content.

Currently, within the developing educational context of Ethiopia, there are very few sophisticated and affordable assistive technologies available to visually impaired students to participate in STEM education. Therefore, the motivation for this research is to develop a framework that can scale and adapt to address these types of contextual challenges while enhancing the access to inclusive and equitable learning opportunities.

In conclusion, the purpose of this research would be to turn inclusive education from a policy idea into a real-life application so that visually impaired students can autonomously have access to, understand, and interact with sight-based educational materials through intelligent assistive devices

1.3. Statement of the Problem

Inclusive education aims to provide equitable learning opportunities for all students, including those who are blind or visually impaired, by ensuring their full participation in mainstream educational settings. However, STEM education relies heavily on visual representations such as diagrams, charts, graphs, illustrations, and schematic models to communicate complex concepts and relationships [22]. Consequently, visually impaired students encounter significant barriers in accessing, understanding, and engaging with visual educational content, limiting their academic participation and achievement compared to their sighted peers [23].

The primary challenge lies in the limitations of existing assistive technologies. Screen readers such as JAWS provide effective access to linear textual information but are unable to interpret complex, non-linear visual content in real time [24]. As such, when visually impaired students access diagramming content (i.e., diagrams that represent electrical circuits or diagrams that represent force vectors), the assistive technology typically provides minimal or non-informative feedback. This lack of information creates a critical data gap that inhibits student understanding of the concepts required and therefore negates the purposes behind inclusive education [25].

Although alternative approaches such as tactile graphics and static image descriptions offer partial support, they are costly to produce, difficult to scale, and often fail to convey the structural, spatial, and semantic relationships embedded within complex diagrams [26].

Recent advances in Artificial Intelligence (AI), particularly in Computer Vision and Natural Language Processing, have demonstrated considerable potential for improving accessibility. However, most existing AI-based accessibility solutions are designed for general-purpose image description and do not provide the detailed, context-aware interpretation required for educational diagrams and STEM learning materials [27][28].

Consequently, visually impaired students continue to experience limited independence, reduced engagement, and unequal access to diagram-based learning resources, particularly in STEM disciplines, resulting in inequitable educational outcomes within inclusive learning environments [29].

To address these challenges, there is a need for an intelligent and integrated assistive technology framework capable of transforming visual educational content into accessible auditory information in real time. Therefore, this study proposes the **JAWS–AI Integrated Assistive Technology Framework (JAIF)**, which combines advanced object detection, Optical Character Recognition (OCR), semantic analysis, and screen-reader integration to interpret STEM visual representations and generate structured, context-aware auditory descriptions through JAWS. By bridging the gap between visual and auditory accessibility, JAIF aims to enhance independent learning, improve access to diagram-based educational content, and strengthen the implementation of inclusive education for visually impaired students through a unified, real-time accessibility solution [30].

1.4. Objective of the study

1.4.1. General Objective

To design and develop a JAWS–AI Integrated Assistive Technology Framework (JAIF) that enhances learning accessibility for visually impaired students by enabling real-time interpretation of visual educational content, particularly STEM diagrams, and delivering structured, context-aware auditory descriptions through seamless integration with screen reader technology.

1.4.2. Specific Objectives

1. To design the system architecture for the JAWS–AI Integrated Assistive Technology Framework (JAIF).
2. To develop a visual detection module using deep learning for detecting and locating elements in STEM diagrams.

3. To implement an optical character recognition (OCR) module to extract textual information from diagrammatic content.
4. To design an NLP-based module to generate structured descriptions from visual and textual features.
5. To create an interface to convert the descriptions into audio output for the JAWS screen reader.
6. To Evaluate the performance of JAIF in terms of detection accuracy, text extraction accuracy, description quality, and processing efficiency.

1.5. Research Questions

1. What system architecture is best suited to build the JAWS-AI Integrated Assistive Technology Framework (JAIF) for reading STEM diagrams?
2. How Can the object detection module accurately identify visual objects in STEM diagrams?
3. How accurately can the OCR module extract text from STEM diagrams?
4. How effectively can the NLP module generate semantic descriptions from detected visual and textual content?
5. How does the JAIF framework connect with JAWS to provide real-time auditory descriptions of STEM diagrams?
6. How to evaluate the overall efficacy of the JAIF framework in terms of detection accuracy, text extraction accuracy, description quality, and processing efficiency?

1.6. Scope and limitation of the Study

The scope of this study was limited to STEM materials with a focus on physics diagrams such as electrical circuit diagrams and free-body diagrams (FBDs), but not other areas of study or pictures that are more complex, or that are not depicted in educational diagrams.

The study is conducted within the context of inclusive education in Ethiopia, particularly in Technical and Vocational Education and Training (TVET) schools and a number of higher educational institutions. The developing of the JAWS–AI Integrated Assistive Technology Framework (JAIF) was informed by the constraints of local technology and educational practice.

The framework has been developed to work with English-language educational content. It does not support multilingual or local language support. The study focuses on design and

implementation of a modular AI pipeline for visual detection, text mining, semantic analysis, and integration with the JAWS screen reader to produce real-time auditory descriptions.

Despite its promising results, the study has several limitations. The relatively small and domain-specific dataset limits generalizability to other diagram types and subject areas. Evaluation was conducted in a controlled environment, which may not fully reflect real classroom conditions. In addition, JAWS integration is constrained by the software's existing design and functionality, limiting further customization. Finally, real-time performance may vary depending on available hardware resources, particularly in low-resource educational settings.

1.7. Significance of the Study

This research makes a significant contribution to the fields of assistive technology (AT), inclusive education, and artificial intelligence (AI) with the introduction of the JAWS-AI Integrated Assistive Technology Framework (JAIF). The intention of the framework is to develop assistive technology that can provide increased access to visually dependent educational material, especially in STEM.

From an educational standpoint, this research provides increased learning access for visually impaired (VI) learners by giving them the ability to interpret diagrams, charts, and other graphical depictions independently. By converting visual information into structured auditory descriptions in real-time, this framework enhances comprehension, engagement, and academic achievement in inclusive learning contexts.

From a technological standpoint, this research expands upon the integration of AI and traditional ATs. By combining computer vision, Optical Character Recognition (OCR), and Natural Language Processing (NLP) through a comprehensive framework, this research expands the capability of screen readers such as JAWS beyond merely providing access to text-based data by enabling the interpretation of complex visual data.

Socially, the proposed framework encourages greater independence and inclusion of VI learners through the reduction in their reliance on external support. This research contributes to reducing

educational inequity and supports global initiatives including the United Nations Sustainable Development Goal 4 (SDG 4) which calls for inclusive and equitable quality education.

The Ethiopian context presents a real-world and adaptable opportunity to address the dearth of access to sophisticated assistive technologies in both inclusive and technical education settings, while also taking into consideration local limitations. Thus, the model developed in this study has the potential to be implemented at scale across technical vocational education and training (TVET) institutions and higher education institutions.

From a research perspective, the study describes a systematic, modular framework that can be used as a foundational model for future development of AI-based assistive technologies. It addresses the disconnect between stand-alone applications of AI and accessible integrated solutions within the broader STEM education context.

From a policy and institutional perspective, the results of the study provide valuable insights for educators, institutions and policymakers about how to create and implement technology-based solutions to better support inclusive education through the use of the framework as a reference for future initiatives for improving accessibility infrastructure and increasing equitable learning opportunities.

1.8. Organization of the Thesis

In this study, there are five chapters dealing with one aspect of this research. The first chapter covers the background, the problem, research objectives, research questions, significance, and limitations of this study. Chapter Two discusses a literature review covering theoretical and empirical perspective on inclusive education, assistive technologies, screen readers, and advances in artificial intelligence, including computer vision and natural language processing, and outlines the conceptual framework for the JAWS–AI Integrated Assistive Technology Framework (JAIF). Chapter Three describes the research design, data collection, system architecture, and model development. Chapter Four presents the system implementation and evaluation of the systems, along with an analysis of performance metric of accuracy and efficiency. Finally, Chapter Five gives the conclusions of the study and suggests further research to improve the JAIF framework.

CHAPTER 2: REVIEW OF RELATED LITERATURE

2.1. Introduction

This Chapter presents the literature on inclusive education, assistive technology for the visually impaired student, and the relevant advancements of artificial intelligence for accessibility via a theoretical and empirical analysis of the two bodies of existing literature, their limitations, and the gaps that exist within the literature that the framework for this study is intended to demonstrate.

Inclusive education is recognized internationally as an essential means for providing equitable learning opportunities to children with disabilities. In this regard, assistive technology is important in aiding visually impaired students to access educational materials and learn in a digital environment. Assistive technology, such as screen readers, has made enormous strides in allowing visually impaired students to access text-based educational materials; however, assistive technology does not adequately address complex visual-based materials that are clearly present in STEM educational materials.

A significant challenge in terms of accessibility arises from increasing use of visual representations (e.g.: diagrams, charts, graphical models) in educational content. Although traditional means of supporting visual content (e.g.: tactile graphics, alternative text descriptions) are useful, they lack real-time interpretation, scalability, and user-interactivity.

Recent advances in AI, such as computer vision, deep learning, and natural language processing (NLP), have demonstrated that the ability to interpret visual information is possible. NLP has been used to generate descriptive outputs from images for various applications. but these technologies are not yet broadly available for use in accessible education and only few of them are used in assistive technology used by visually impaired students.

This chapter therefore review key studies that deal with (1) inclusive education and accessibility challenges, (2) current assistive technology and their limitations, and (3) interpreting images using AI. Based on this review, the chapter identifies gaps in existing solutions and proposes an explanation for the JAWS–AI Integrated Assistive Technology Framework (JAIF).

2.2. Theoretical Literature Review

2.2.1. Overview of Deep Learning and Object Detection

Deep Learning, a part of Artificial Intelligence (AI) has further improved the ability of machines to analyze and interpret visual information by automatically extracting hierarchical features from large-scale data. This makes it an excellent tool in image analysis, particularly where it is critical to interpret structured visual information accurately. Object detection is important in image analysis both for localization (which tells us where an object is located in an image) and classification (which tells us what objects belong to which categories)[31][32].

Earlier object detection methods employed multiple stages of feature extraction, region proposal, and classification. While effective, these approaches are computationally intensive and not well suited to real-time applications. To remedy these shortcomings, modern detectors based on deep learning have used single-stage detector architectures that do both detection and classification simultaneously, increasing processing speed and efficiency[33].

The You Only Look Once (YOLO) family of algorithms represents one of the most promising approaches in that it focuses on object detection as a unified regression problem and enables real-time object detection with competitive accuracy[34].Recent advances in object detection models also emphasize the trade-off between speed and accuracy, often measured by mean Average Precision (mAP).

These advances are especially useful for interpreting structured and detailed visuals such as STEM diagrams where a precise detection of small and complex objects is required. For example, we consider modern deep learning-based object detection as a technical foundation for real-time visual interpretation in assistive systems such as the proposed JAWS–AI Integrated Assistive Technology Framework (JAIF)[19].

2.2.2. Assistive Technology in Inclusive Education

Inclusive education promotes an educational system that allows all students, including the visually impaired, to participate in the same classroom content. For the visually impaired Students, this can be a challenge, especially in Science, Technology, Engineering, and Math (STEM), where diagrams, graphs, and other visual representations are essential for understanding complex

concepts and relationships[35].The accessibility of such materials hinders comprehension, participation, and academic performance..

Assistive technologies address these issues by converting images to alternative formats, such as audio or tactile. Assistive technologies such as screen readers and tactile graphics act as a bridge between the visual content and non-visual perception to help users access and interpret graphical information. But, typical assistive solutions can only interpret the visual structure of diagrams, especially in the context of real-time learning[36]. As a result, aids may only partially help access STEM visual content.

The theoretical foundation of this study are the Universal Design for Learning (UDL) principles, which advocate flexible instructional practices to meet the needs of a variety of learners. UDL demonstrates the importance of multiple representations of information in order to present it in an accessible manner for all learners[37]. Artificial Intelligence used to interpret visual information automatically fits into the UDL principles by converting diagrams into auditory descriptions in real-time; this supports inclusion and accessibility in the learning environment and forms the basis for the JAWS-AI Integrated Assistive Technology Framework (JAIF) [38].

2.2.3. Screen Readers and Digital Accessibility (JAWS)

Screen readers are assistive software applications which render digital content onto a screen in a non-visual format such as synthesized speech or Braille for visually impaired users. JAWS (Job Access With Speech) is among the most popular screen readers for Windows. This software works well with structured text such as documents, web pages, and user interface elements such as HTML[39].

JAWS is a great tool for text-based information, but it cannot take in non-textual visual content (such as images, diagrams, and other visual representations of STEM content), which causes the system to give only limited or no information[40].This creates a barrier for accessibility when visual information is needed.

Recent advances in Artificial Intelligence have been used to automatically interpret visual information by extracting structure, text, and semantic information from images and transforming it into structured representations. When embedded with screen readers via APIs, this can allow the

user to access visual content beyond traditional text processing, using visual information to form context-sensitive descriptions via auditory representations. This integration extends the functionality of traditional screen readers and is one of the key elements of the proposed JAWS–AI Integrated Assistive Technology Framework (JAIF), providing real time access to diagrams[41].

2.3. Empirical Literature Review

This section reviews the empirical studies published on the global and local levels regarding AI-based assistive technologies that improve accessibility of STEM education for visually impaired students. It reviews existing approaches and evaluates their success in solving the accessibility challenges of visual content.

Studies are grouped by technology, applications, and context to enable a clear understanding of the evolution of assistive technologies, such as artificial intelligence such as computer vision and natural language processing. A focus on how these technologies help interpret images and help access is highlighted.

Lastly, this section reviews strengths and limitations of prior studies in their ability to interpret complex diagrams, respond to a graphical image in real time, and integrate into assistive technologies. Despite significant progress, many solutions are limited and address isolated problems without offering a fully-integrated system.

These limitations expose a critical research gap, that is, there is no unified, scalable framework for integrating visual information from AI with assistive technology and application in real time. This gap is the foundation of the JAWS–AI Integrated Assistive Technology Framework (JAIF), which was help STEM users gain a structured, context-aware hearing access to STEM visuals.

2.3.1. Evolution of YOLO Algorithms in Assistive Contexts

The use of real-time object detection in accessibility has increased with the development of the YOLO (You Only Look Once) architecture. Early research, such as Srivastava et al. (2020), used YOLOv3 to identify simple electrical components and demonstrates the practicality of deep learning for assistive visual interpretation. These models, but, were relatively slow and inefficient

in real-time assistive scenarios[42]. Subsequent work has improved on performance and applicability. For example, Zhang et al. (2024) used YOLOv5 to detect general classroom objects in order to aid students' awareness and navigation in classrooms. Though, these models were made for general objects and were not designed to accurately detect fine-grained STEM symbols[43]. Recent work has focused on improving accuracy and computational efficiency. Newer versions such as YOLOv8 and YOLOv10 are faster and more precise, resulting in more accurate detection of real-time objects in assistive applications[44][45]. Small, dense, and complex elements from educational diagrams remain challenging to accurately detect.

These shortcomings show the need for specialized detection for STEM diagram reading. The present work adopts a modern YOLO model to help detect fine-grained images in an accurate and real-time way within the proposed JAWS-AI Integrated Assistive Technology Framework (JAIF).

2.3.2. AI for STEM-Specific Recognition and Diagramming

Beyond detection of objects in general, Artificial Intelligence has also been used to decode scientific knowledge. For instance, Deshpande (2024) converted mathematical equations into machine-readable formats using Convolutional Neural Networks (CNNs) that provided structure to an equation in a simple form. But, this work was only applicable to images processed in static mode and was not interactive[46].

Park, Sung, and Kim (2025) addressed this need with a mobile-based circuit recognition system, which offers immediate feedback to visually impaired Students. While this approach improved interaction, it was limited to a certain type of diagram, and not universally applicable across STEM subjects[47].

In addition, Yan et al. (2025) developed Transformer models to generate Alternative Text (Alt-Text) descriptions of scientific graphs, bringing semantic understanding to the accessibility issue. While they generate meaningful descriptions, these models do not integrate with assistive technologies for real-time interaction[48].

Overall, these studies show significant gains in STEM recognition and understanding of meaning but are limited in scope, tackling isolated tasks, such as detection, recognition, or description,

without providing a single and interactive solution. This lack of integration with assistive technologies such as screen readers highlights a critical gap, and indicates the need for a whole-of-system approach that combined accurate diagram interpretation with auditory feedback, such as the JAWS–AI Integrated Assistive Technology Framework (JAIF).

2.3.3. Integration with Screen Readers and User Experience

Assistive technologies are evaluated according to how seamlessly they integrate with existing accessibility tools. In this vein, Gao, Manduchi, and Legge (2025) study how optical character recognition (OCR) may be integrated with the JAWS screen reader for accessibility improvements to text. Although their approach is effective for reading text, it is not particularly useful for reading complex visual elements such as diagrams and graphs[49].

Recent work highlights the need to further integrate assistive technology into other processing methods and extend assistive technologies beyond their traditional, just functional capabilities. Some assistive technology applications use artificial intelligence for text-encoded or -described content, yet are often not fully integrated with screen readers for real-time interaction, leading to fragmented access solutions.

These deficiencies require a more integrated approach to AI-based visual interpretation with screen readers. The aim of this research is to integrate automated visual analysis with the JAWS interface to convert complex visual content into structured auditory input. This integration could allow users to access diagrams in real time and is part of the proposed JAWS–AI Integrated Assistive Technology Framework (JAIF).

2.3.4. Local Context and Accessibility Barriers in Ethiopia

In Ethiopia, access for the visually impaired Students is constrained by the poor state of technology, resources, and aids available. They have difficulty with the STEM learning material, where drawings, diagrams, or other visual representations are important.

Data from Ethiopian higher education institutions indicate that visually impaired students may be discouraged from choosing STEM fields because of a lack of assistive technology and institutional

support for the visually impaired Students. The educational environments do not accommodate accessibility needs for visual content[37].

These results highlight a gap in the development and implementation of accessibility solutions based on AI around the world. To close this gap, we need to find an adapted and context-sensitive solution relevant to local technologies and educational practices, especially within Technical and Vocational Education and Training (TVET) institutions[50].

Overall, despite AI-based assistive technologies having made significant advances, many currently available systems are domain-specific, unintegrated in the real-time, and may not fully connect with screen readers. As mentioned above, when coupled to the challenges faced by Ethiopians, these findings suggest that an integrated and adaptable solution needs to be developed to help students with visual educational content understand the content. Such a solution would be the basis for the JAWS–AI Integrated Assistive Technology Framework (JAIF)

Table 1 Summary of Reviewed Literature

Author(s) & Year	**Focus Area**	**Method / Technology**	**Key Contribution**	**Limitation**	**Relevance to This Study (JAIF)**
Srinivasan et al. (2020)	Assistive object detection	YOLOv3	Demonstrated feasibility of real-time object detection for visually impaired Students	High latency; limited accuracy for complex scenes	Establishes baseline for AI-based assistive detection
Zhang et al. (2024) [43]	Classroom accessibility	YOLOv5 + embedded systems	Improved navigation and environmental awareness	Not suitable for structured STEM diagrams	Highlights need for domain-specific detection models
Wang et al. (2024) [45]	Advanced object detection	YOLOv10	Enhanced detection speed and efficiency	Difficulty detecting small/dense objects	Motivates improved detection for diagram components

Hasan et al. (2025) [34]	Object detection evolution	YOLO architectures review	Provides comprehensive analysis of detection improvements	Lacks application to assistive systems	Supports selection of advanced YOLO-based models
Deshpande (2024) [46]	Mathematical recognition	CNN-based model	Converts equations into structured digital form	Static processing; no real-time capability	Highlights need for dynamic and real-time interpretation
Park et al. (2025) [47]	Real-time assistive systems	Mobile-based recognition	Enables real-time interaction for visually impaired users	Limited to specific use-case (gesture/circuit)	Shows importance of real-time processing
Yan et al. (2025) [48]	Graph accessibility	Transformer-based Alt-Text generation	Generates semantic descriptions of charts	No integration with assistive tools	Supports need for NLP-based semantic interpretation
Gao et al. (2025) [49]	Text accessibility	OCR + screen reader integration	Improves text extraction for visually impaired users	Cannot interpret diagrams or spatial relationships	Justifies integration of OCR in JAIF
Muhsin et al. (2024) [36]	Assistive technologies review	Survey of AT tools	Identifies limitations of traditional assistive tools	Lack of real-time and intelligent interpretation	Establishes gap in existing assistive systems
Aleme (2023) [50]	Ethiopian accessibility context	Empirical study	Identifies barriers in STEM accessibility	No technical solution proposed	Justifies need for localized AI framework

2.4. Related Works

Srinivasan et al. (2020) proposed a deep learning-based object recognition system based on YOLO-type architectures that could assist visually impaired Students identify objects in their environment. The system was effective in general-purpose object recognition, but it was not designed to interpret structured diagram elements such as electrical or physics symbols, and its performance was low when applied to visually complex content[42].

For the present work, we employ YOLOv12, an advanced object detection model optimized for speedy and precise detection. Unlike the earlier models, it is specifically designed to deal with the

dense and complex symbolic structures in physics diagrams in TVET contexts. This improves the detection accuracy and interpretation of diagrams in real-time, thus overcoming the limitations in previous work.

Li et al (2022) proposed a Counting-Aware Network (CAN) to handle the structural complexity of handwritten mathematical expressions. The model is capable of parsing multi-line formulas via symbol counting, with a better understanding of the structural relationships present in scientific notations. It is only designed for high-resolution images and cannot be processed in real time[51].

As a result, the present study uses YOLOv12 to detect and analyze visual objects in real-time in a dynamic learning environment such as physics labs. It also integrates the output from the sensor with the JAWS screen reader to provide immediate feedback via sound, moving from a purely static detection process into a more interactive and accessible learning tool for visually impaired students.

Arifando et al. (2023) proposed a novel YOLOv5-based object detection model to assist visually impaired users in recognizing objects in their environment. The system detected objects more accurately and allowed the user to walk autonomously. But, the model does not capture specific visual content such as physics diagrams or scientific symbols[52].

Our study has employed a YOLOv12 model trained on science and diagrams to enhance access to academically relevant information. It correctly recognizes structured and complex STEM visuals, which is a potential issue in assistive technology that facilitates access to STEM education.

In developing a mobile real-time object recognition system for visually impaired users, Kim et al. (2023) demonstrated the feasibility of using deep learning models on smartphones with auditory feedback within a few seconds. But, while this approach is portable and offers real-time performance, it has no mechanism for providing coherent description of the domain, especially with STEM-related content[53].

In contrast, the present study proposes a unified processing pipeline that integrates detection of objects via YOLOv12, OCR, and semantic interpretation, to produce structured and context-

sensitive descriptions via the JAWS screen reader, which was help access and understanding of physics diagrams in schools.

Aleme (2023) analyzed the adoption of inclusive education for persons with disabilities in Ethiopian higher education, and identified institutions, technologies, and teachers that made it difficult for visually impaired students to participate and access instruction. The findings point out challenges in STEM courses where visual content is an essential ingredient, but do not propose a practical solution[50].

In contrast, the current study builds an AI-assisted framework based on the YOLOv12 architecture that allows students to interpret STEM diagrams in real time and receive feedback through screen readers, which was allow inclusive teaching and learning in Ethiopian TVET and Higher education.

2.5. Research Gap

Though recent research supports the use of deep learning and assistive technology to support visually impaired Students, there are limitations related to inclusive STEM education. Existing YOLO-based strategies recognize objects and navigate towards them; but it is problematic to interpret dense and structured diagrams, such as those used in physics or electrical symbols. Similarly, mathematical expression recognition methods can parse complex structures, but only from images and do not address interaction or feedback appropriate for classrooms[51] .

Although aids on mobile devices provide real-time information and portability, they do not provide a way to create coherent, domain-specific descriptions of audio needed in learning STEM[53]. Studies in Ethiopia have identified institutional, technological, and educational barriers to visual impairments but have not identified practical technological solutions[50]

These limitations indicate the absence of an integrated, real-time, and context-aware assistive system that combines diagram interpretation with screen reader technologies. In this thesis, we propose a YOLOv12-based system that combines object detection, OCR, and semantic processing with the JAWS screen reader. The proposed system is designed to interpret physics diagrams in real time and meets the needs of Ethiopian TVET and inclusive education.

Table 2.Comparative Analysis of Related Works

Study	Focus Area	Method/Technology	Key Contribution	Limitation	Gap Addressed by This Study
Srinivasan et al. (2020)[42]	Object recognition for accessibility	YOLO-based models	Demonstrated feasibility of object detection for visually impaired users	Limited to general objects; poor performance on complex diagrams	Introduces YOLOv12 for accurate detection of dense STEM diagrams
Li et al. (2022)[51]	Mathematical expression recognition	Counting-Aware Network (CAN), CNN	Improved parsing of multi-line mathematical expressions	Restricted to static images; no real-time interaction or audio output	Enables real-time processing with auditory feedback via JAWS
Arifando et al. (2023)[52]	Assistive object detection	YOLOv5-based model	Enhanced detection accuracy and user autonomy	Focused on navigation; lacks STEM-specific interpretation	Develops domain-specific detection for physics diagrams
Kim et al. (2023)[53]	Mobile assistive system	Deep learning on smartphones	Demonstrated real-time detection and portability	No integration for structured, context-aware descriptions	Integrates YOLOv12 + OCR + semantic processing for coherent output
Aleme (2023) [50]	Inclusive education (Ethiopia)	Empirical accessibility study	Identified institutional and technological barriers	No technical solution proposed	Provides localized AI-based framework for Ethiopian higher education & TVET context

2.6. Conceptual Framework

The conceptual framework of this study illustrates the logical progression of visual educational content into accessible auditory information, the JAWS-AI Integrated Assistive Technology Framework (JAIF). In particular, it demonstrates visual input, intelligent processing, and accessible output, as well as a feedback loop for continuous improvement[54]. This framework

attempts to address the limitations of traditional assistive technology, by offering a more unified, real-time approach to interpreting complex visual information, especially in STEM education [55].

Initially, we start with visual educational inputs as the main source of information, i.e., diagrams, charts, graphs, and other graphical representations found in digital textbooks, in scanned learning materials, or online learning platforms. Since these types of information come in many forms, we preprocess the data to make it usable[56]. Our preprocessing processes include resizing, normalizing, noise reduction, and increasing contrast to standardize and standardize the visual content for later analysis

The next step is to employ an AI-driven process as the transformation stage. At this stage, different intelligent modules interpret the visual input. Object detection techniques such as the YOLOv12 model are used to identify and locate objects such as shapes, arrows, symbols, and building elements within a diagram. Optical Character Recognition is used to extract text from the visual content such as labels, annotations, and numbers[57]. Both sets of outputs are then combined and processed using Natural Language Processing. These techniques interpret the relationship between detected elements and produce a context-aware representation of the diagram. This step transforms the visual input into structured information[58].

The content is then translated to an accessible form for visually impaired Students. First, the structure of the semantic representation is converted into a textual description that accurately reflects the meaning and relationships of the diagram[59]. Next, the output of the text is transformed into audio which can be delivered to a screen reader like JAWS in real time, allowing the user to access and understand the visual content through spoken words. This change from the visual input to the audio output is the goal of the framework in providing a more accessible and inclusive learning environment[60].

In addition to the transformation, the framework includes a feedback and optimization component that allows it to be continually improved upon. The feedback consists of system-level feedback and the user experience, in terms of accuracy, response time, usability, and clarity of generated descriptions[61]. The feedback provides useful feedback to improve the performance of the AI models, semantic interpretation, and responsiveness of the system. The framework is an iterative and adaptive process.

Overall, this conceptual framework represents a continuum of change as visual information is interpreted, and delivered as accessible auditory information. This conceptual model includes computer vision, Optical Character Recognition, Natural Language Processing, and assistive technology to provide a structured and flexible solution for visual content accessibility[62]. It shows how Artificial Intelligence can further enhance the capabilities of traditional screen readers like JAWS to provide more inclusive, independent, and effective learning experiences for visually impaired students.

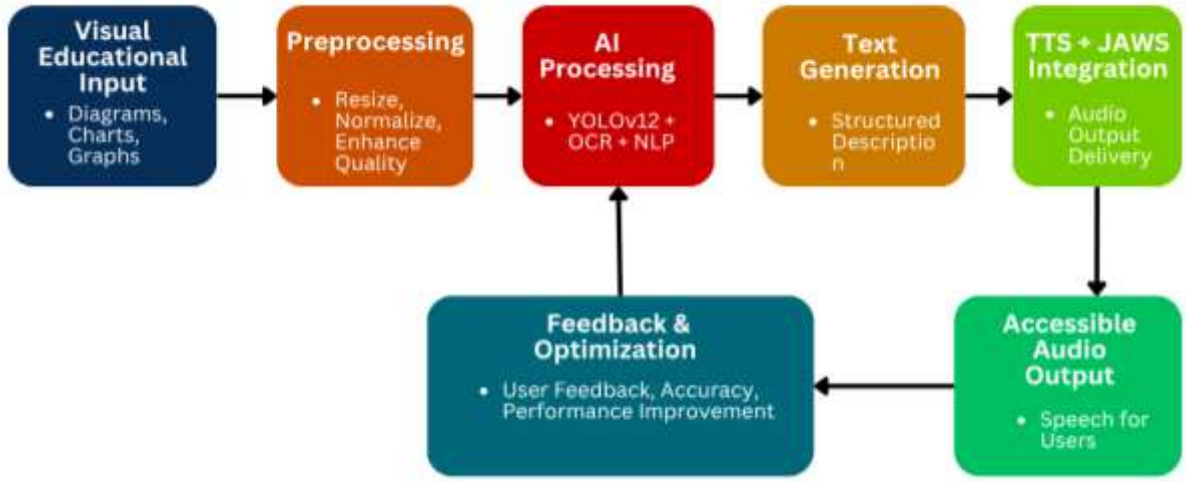


Figure 1.Conceptual Framework

Figure 1. presents the conceptual framework of the proposed JAWS-AI integrated assistive technology system. This framework illustrates the interaction between content that is visually inaccessible and its audio equivalent; it extends the scope of screen readers like JAWS as well as facilitating more inclusive and equitable learning environments.

CHAPTER 3: RESEARCH METHODOLOGIES

3.1. Introduction

This chapter presents the process for designing, developing, and testing the proposed JAWS AI Integrated Assistive Technology Framework (JAIF). Building upon the conceptual and theoretical components presented in the last chapter, this section establishes a system that can address accessibility issues that visually impaired children and adults experience when reading visual educational content.

The study approaches the task through a system-oriented approach in which diagrams and graphs, particularly in STEM areas, are used as primary inputs. These inputs are converted to an accessible text using a sequence of steps: image preprocessing, object recognition, optical character recognition, language comprehension, and text-to-speech translation. All this is combined with JAWS' screen reader assistive technologies.

The research uses Design Science Research (DSR) methodology to design the JAIF framework as a functional artifact, and focuses on an experimental design of a controlled experiment designed to assess the system's performance under preset, repeatable conditions.

The performance of the system, which consists of image preprocessing, object detection, optical character recognition, semantic interpretation, and text-to-speech is evaluated using standardized data sets and simulated input variation. Performance is measured with objective measures such as precision, recall, F1 score, mean Average Precision (mAP), Character Error Rate (CER), Word Error Rate (WER), and processing latency. A baseline comparison is used to measure performance improvements versus baseline. This chapter describes more fully the design, system architecture, preparation of data sets, experiment setup, and assessment measures for the reliability and validity of the study.

3.2. Research Design and Approach

This research used Design Science Research (DSR) approach to develop and evaluate new technology artifacts that solve real-world problems. DSR focuses on creating a logical and functional system, which is then evaluated using measurable performance criteria.

The study design, develop, and evaluate an AI-integrated assistive technology framework (JAIF) that facilitates the automatic interpretation of visual educational content. We first explore the problem, create the artifact, and then evaluate its performance.

At the Architecture level, this thesis defines a modular system architecture made up of interconnected components such as object detection, optical character recognition (OCR), natural language processing (NLP), and text-to-speech (TTS). Each module is built to perform a specific task in the overall pipeline, offering flexibility, and efficient processing of visual data.

The framework is implemented as a data-driven AI pipeline that processes visual inputs such as diagrams and graphic content through detection, extraction, interpretation, and audio generation. This enables processing in real-time visual information into machine-interpretable representations.

Evaluation is based on quantitative metrics, such as detection accuracy (mAP), text extraction accuracy (CER, WER) and description quality (BLEU, ROUGE). Thus, the effectiveness of the proposed system is evaluated objectively based on the algorithm rather than by the user interaction.

Overall, the research design is systems-oriented and experimental, aiming to build and test an AI-based system. Its main objectives are technical performance, architecture, and real-time processing; that way, it addresses the limitations of existing assistive technology for complex visual content.

3.3. Model Architecture

The JAIF system consists of a series of modular processing steps to turn visual and written instruction into auditory information. We use a modular approach over an end-to-end system for transparency, flexibility, and compatibility with assistive technologies like JAWS. Each

component of the system is used independently, while contributing to the overall functionality of the system, so that each component can scale up and down as necessary.

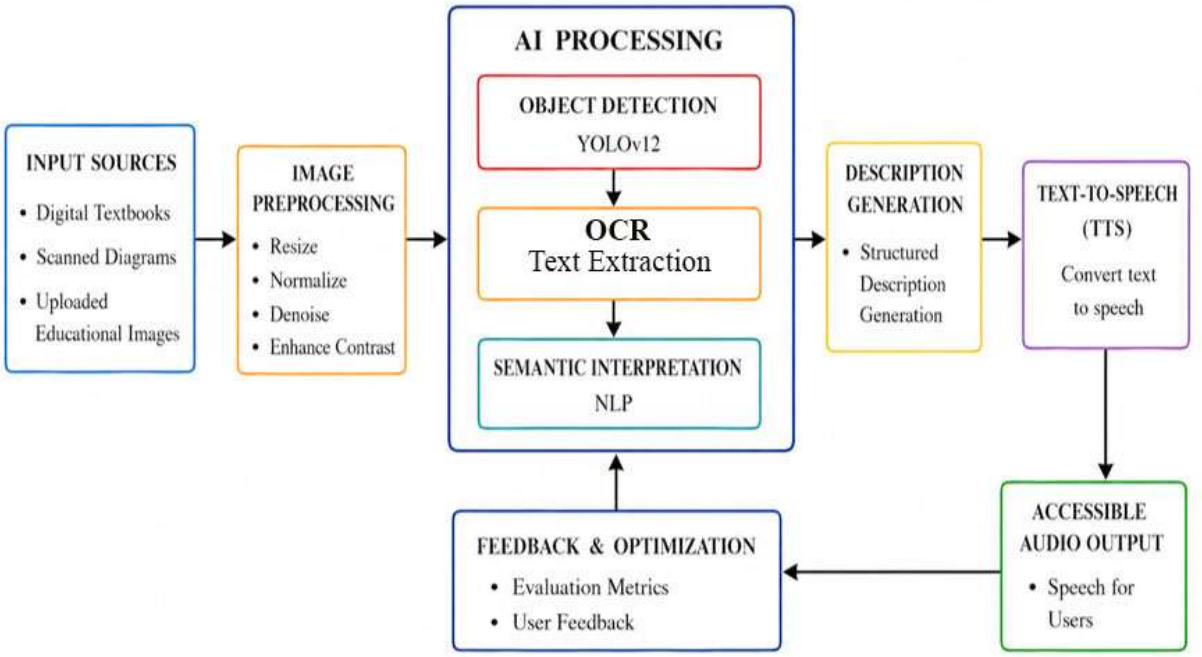


Figure 2.Proposed Model of Architecture for JAIF prototype

The system begins with obtaining visual input in the form of diagrams, charts, or other graphic elements from digital or scanned sources. Preprocessing includes resizing, normalization, noise reduction, and contrast enhancement to improve image quality and input consistency.

After preprocessing, the system performs AI-based analysis where object detection is used to identify structures in the diagram such as shapes, arrows, and symbols. Optical character recognition is used to extract textual information embedded in the visual content such as labels and annotations. The results of these two processes are combined using natural language processing techniques to create coherent and contextual descriptions of the diagram.

The resulting textual descriptions are translated into speech through a text-to-speech module, allowing the user to hear the text. The wording and textual descriptions are output to the JAWS screen reader, allowing the visually impaired Students to interact with the information at any time.

While a fully multimodal model may work well, the need for a pipeline-based structure is not justified by the need for interpreting, controlling, and linking with existing assistive technologies. In an educational setting, particularly one with limited resources, it is imperative that the system can be easily adapted and adapted in order to make small changes and debugging easier. This way, the system can be deployed and maintained in a way that meets the accessibility needs of the user.

3.4. Model Selection and Justification

Models and technologies used within the JAIF framework are chosen based on the ability of the model or technology to be used in real-time, accuracy in the specific domain, and compatibility with assistive technologies. Each part of the system is selected for its ability to process STEM diagrams in a resource-constrained setting.

3.4.1. Object Detection Model (YOLOv12)

The YOLOv12 model is selected for object detection because of its high precision in real-time detection in complex visual environments. While two-stage detectors like Faster R-CNN focus on accuracy at the expense of speed, YOLOv12 is a one-stage detector with considerably lower latency than those using two-stage detectors.

Comparing with versions like YOLOv5 and YOLOv8, YOLOv12 has been enhanced with better feature extraction and attention-based enhancements that allow it to find small, dense, and overlapping objects in STEM diagrams such as arrows, circuit symbols, and coordinate axis. This makes YOLOv12 an ideal solution for the proposed system, where precision and quick response are important

3.4.2. Optical Character Recognition (OCR)

Optical Character Recognition (OCR) is used to decipher the textual information encoded in diagrams (such as labels, annotations, and values). Two OCR tools are considered: Tesseract and PaddleOCR.

Tesseract is chosen because it is lightweight and open-source, and it is easy to integrate so it is ideal for use in low-resource environments. PaddleOCR is selected because it has the best performance in handling complex layouts and multilingual text. OCR is essential as a way of

understanding the meaning of a diagram, not only in its structure, but also in the extraction of textual information supporting the meaning.

3.4.3. Natural Language Processing (NLP) Model

In this thesis, we apply a Transformer-based NLP approach to describe diagrams in a structured and context-sensitive way. Unlike rule-based models that are not capable of handling complex relationships, Transformer models can capture context-dependent relationships between visual components and textual annotations.

Models such as BERT, T5 or GPT can be used for this task because they allow for coherent and semantically rich descriptions, moving beyond the mere listing of objects to a meaningful explanation of diagram structure and relationships.

3.4.4. Text-to-Speech (TTS) Model

Text-to-Speech is responsible for translating the generated descriptions into speech. For beginners, it is best to use lightweight TTS solutions such as gTTS because they are both light and low computation intensive. Advanced neural TTS systems like Tacotron 2 or FastSpeech 2 produce more accurate and natural speech. TTS should be selected based on the trade-off between speech quality and latency, and with the priority being for responsiveness in real time for educational use.

3.4.5. JAWS Integration Layer

JAWS is chosen as the first accessible interface because of its broad use in schools and professional settings. JAWS scripting and accessibility APIs are used to integrate into the software for automatic delivery of the resulting auditory descriptions. By contrast to other screen readers like NVDA, JAWS is much more compatible with enterprise systems and provides structured navigation, making it suitable for use in educational settings.

3.5. Data Collection

Data collection in this study is limited to the technical dataset of STEM-based diagrammatic images used in development of the proposed JAWS–AI Integrated Assistive Technology Framework (JAIF). The dataset contains exactly 200 STEM diagram images from reliable sources, such as digital textbooks, open educational resources (OER), and publicly accessible academic repositories. The data set is focused on physics diagrams (e.g., electrical circuit diagrams, free-

body diagrams, and basic graph representations) because they are highly structured and relevant to the study objectives.

A random selection of 200 images ensures diversity in structure, layout, labeling, and visual complexity. The controlled size of this data set allows for prototype-level system development and testing within Design Science Research programs with the aim of validating the system rather than large-scale generalization.

All images were converted to JPEG and PNG files for use with the supervised learning pipeline. Images were manually annotated with Roboflow. Icons associated with the axis, arrows, symbols, objects and text were marked with bounding boxes. Annotations were carried out directly with the object detection and OCR functions of the system.

Overall, the way data is collected is strictly technical, controlled, and data driven in accordance with the end goals of the study, which are to develop an automated framework for interpreting STEM diagrams without using data about users.

3.6. Data Collection Instruments

The means used in this study to collect data are tools and software used to collect, manage, annotate, and validate the STEM image dataset. The primary tool used in preparing the dataset is Roboflow, which allows for the manual annotation of the images and in managing the dataset. Roboflow was label the STEM diagrams from the images by defining boxes around the elements of the diagram such as shapes, arrows, symbols, and text. The dataset can also be reorganized, saved in versions, and exported in YOLO format so that the data is usable by the object detection model.

In addition to the tools for annotation and image management, digital resources such as open educational repositories and image extraction tools from textbooks provide great sources for high-quality and diverse STEM diagram images. To ensure quality and consistency, we apply a structure in the annotation process of the data during the labeling step such as predefined class labels, consistent placement of bounding boxes, and verification of the annotated images. All annotated data are reviewed to check for accuracy in labeling and consistency across images.

Further file handling and preprocessing tools such as image format converters and basic image processing utilities are used to convert images into standard formats (JPEG/PNG) and prepare the dataset for model training. The tools used in data collection are entirely technical and software-based and allow for a controlled and systematic process of preparing the dataset for use in evaluating and developing the JAIF framework.

3.7. Data Preprocessing and Preparation

Preparing and preprocessing data are critical parts of making sure that the JAIF framework is reliable and efficient. Preparation helps to control and improve the quality of images before they are sent to the AI processing modules.

First, the visual data is cleaned and sorted to eliminate duplicates, samples not complete, and images with low quality that might affect the object detection model. Images are then converted into JPEG and PNG to fit the processing pipeline. In order to improve accuracy, images are resized to a uniform resolution for faster processing and consistent input to the object detection model.

Preprocessing is applied to improve image clarity and feature visibility. It includes normalizing the pixel intensity, noise reduction to remove distortions during scanning or compression, and contrast enhancement to improve the visibility of diagram components such as lines, symbols, and text. These steps are critical for improving object detection accuracy and OCR performance.

After the image data has been pre-processed, we structure the data for supervised learning. Annotated images are validated to check the labeling consistency and accuracy of the image, particularly for components such as shapes, arrows, symbols and text. This is important to ensure that our model does not have training errors and that the model generalizes.

Afin to support effective model training and testing, the dataset is separated into training, validation, and testing subsets. The training set is used to train the object detection and OCR models. The validation set is used to improve model parameters and to avoid overfitting. Finally, the testing set is used to test the performance of the system under unseen conditions.

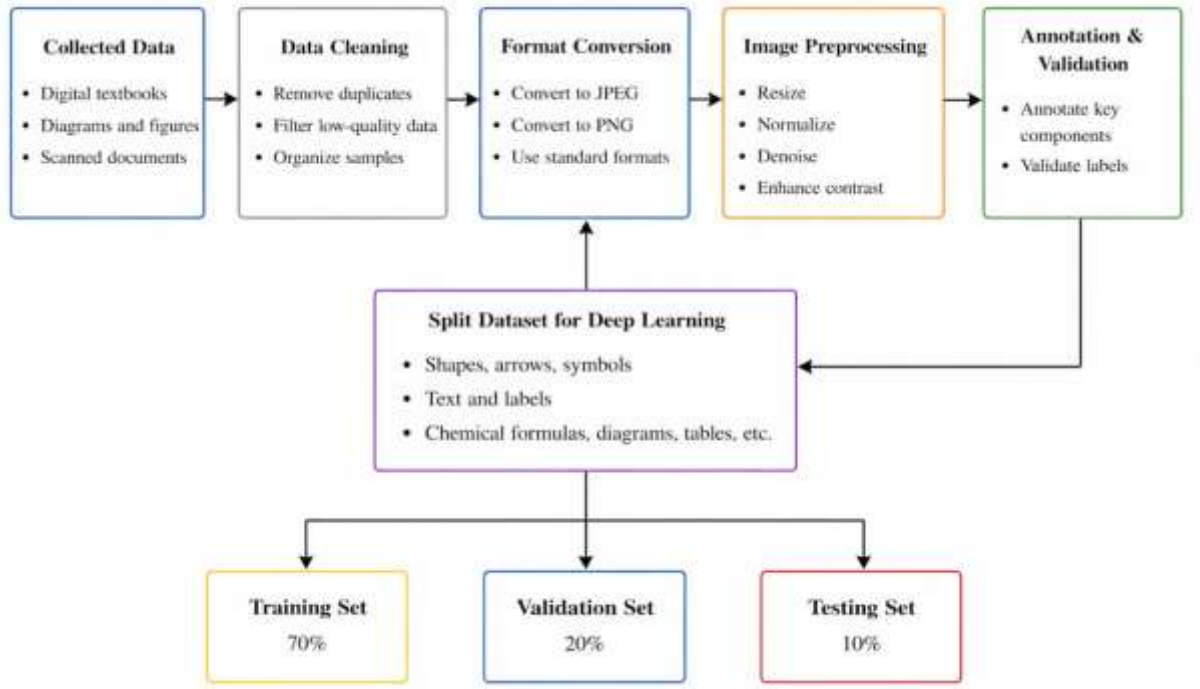


Figure 3.Data Preprocessing and Preparation

Figure 3 shows the data preprocessing and preparation process used in the proposed JAIF framework. Images are cleaned, resized, normalized, and their noise and contrast are enhanced to improve image quality and ensure uniformity before putting the AI models through training.

3.8. Data Split

To increase the confidence in our models (reliability, robustness, generalizability), we break the data into three subsets: training (70%), validation (20%), and testing (10%). This allows the model to learn, be tuned, and be evaluated objectively.

The training set was used to train the model to recognize objects, OCR (Optical Character Recognition) and translate to meaning. The validation set was used during model development to monitor performance, adjust parameters, and prevent overfitting. The test set is completely separate from the training and validation sets and was not used during training or validation. The test set was used at the final evaluation to observe how well the model performed on unseen data.

To ensure data integrity and avoid leakage, splits were performed prior to training of the model so no overlapping samples were present between set 2 and set 3. By splitting the data prior to training, the model is more likely to generalize to other types of diagrams and real-world learning problems.

3.9. Model Training and Feature Extraction

The processing of model training and features is an important step in the JAIF system, enabling us to interpret visual educational content accurately. This process is closely linked to the AI processing, where visual and text information are extracted and transformed into meaningful representations.

Feature extraction at two levels: visual and textual. Visual features are extracted using the object detection model YOLOv12. The model detects the structural components in diagrams, such as shapes, arrows, symbols, and graphic elements. The model learns spatial relationships through supervised training on annotation datasets, thus being able to detect many objects in a visual environment efficiently.

Parallel to the visual features, text is extracted by optical character recognition (OCR) such as labels, annotations, or numerical values. The textual information can provide background information to the visual features detected by the object detection model.

As a result, the outputs of object detection and OCR are combined via fusion, and then passed to a Transformer-based natural language processing model for semantic interpretation. This model analyzes relationships between images and text, and generates structured, context-aware descriptions of the diagrams it reads.

Model training is performed using annotated datasets prepared in the data preparation phase. To allow for more objective evaluation and to minimize overfitting of the models, the dataset is divided into training, validation, and testing subsets. The model YOLOv12 is trained with supervised learning with labeled boxes, while the OCR and NLP components are trained using models adapted to educational diagrams.

Overall, the combination of object detection, OCR, and NLP enables the detachment of features from diagrams and interpretation of these features so that JAIF can generate accurate and meaningful descriptions that are accessible to visually impaired Students

3.10. Libraries and Software Tools

The proposed JAIF system is implemented using various languages, libraries and software tools that support computer vision, deep learning, natural language processing, and assistive technology. It has been developed using Python due to its flexibility, scalability, and extensive ecosystem of artificial intelligence applications that allow seamless integration of different modules of the processing pipeline.

In the processing of visual data, the OpenCV library was used to do some preliminary work such as resizing, noise reduction, and contrast enhancement to ensure that the visual data we are processing is optimized for analysis. Object detection was done using the YOLOv12 model of the PyTorch deep learning framework, which is designed to train and infer the correct representation of shapes, symbols, and graphic elements in diagrams.

Text within diagrams is extracted by using Optical Character Recognition (OCR) tools such as Tesseract and PaddleOCR to identify labels, annotations, and numerical values providing contextual information on top of the visual features. Semantic interpretation is accomplished through the use of Transformer-based models implemented in Hugging Face Transformers library, which integrates outputs from object detection and OCR modules to provide a structured and context-aware description.

Integration with assistive technology is accomplished by JAWS scripting (JSL) which allows the interaction between the system and the JAWS screen reader, allowing textual descriptions to be delivered as speech in real-time, allowing users with visual impairments to access them.

These libraries and tools integrate seamlessly into the JAIF processing workflow from image preparation and feature extraction to semantic processing and output. This combination guarantees the highest system performance, modularity, and compatibility with assistive technologies.

3.11. Performance Evaluation Method

The performance of the proposed JAWS–AI Integrated Assistive Technology Framework (JAIF) is evaluated using technical performance metrics to assess the system's effectiveness in processing and interpreting visual educational content. The evaluation is conducted under controlled, experimental conditions to evaluate accuracy, efficiency, and reliability of the proposed system.

Technically, the object detection component is assessed using standard parameters such as precision, recall, and mean Average Precision (mAP) in order to measure the accuracy and reliability of the detection and location of diagram elements including shapes, symbols, and graph structures.

Performance of the Optical Character Recognition (OCR) module is measured by Character Error Rate (CER) and Word Error Rate (WER). These measures measure whether the system can accurately extract text (labels, annotations, and numeric values) from diagrams.

For the Natural Language Processing (NLP) component, the quality of generated descriptions is assessed by the BLEU and ROUGE scores, which measure the similarity between system-generated outputs and reference descriptions. These scores ensure that generated descriptions are coherent and semantically meaningful.

Efficiency is also measured by measuring processing time and response latency. These factors reflect the ability of the system to perform in real-time assistance applications.

In addition, the robustness of the framework can be derived by testing the system's performance under different input conditions, such as image quality and diagram complexity, to ensure that the system performs consistently across various inputs.

Overall, the integration of these technical evaluation measures allows for a full assessment of JAIF. The system is balanced between accuracy, efficiency, and reliability, meaning it is effective as an assistive technology in enabling access to visual educational content.

Precision: calculated as the ratio of true positive detections to the total number of predicted positives, which includes both correct detections and incorrect detections. It can be represented by the following formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{F}P}$$

Recall: Recall is calculated as the ratio of true positives to the sum of true positives and false negatives. It can be represented by the below

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Confusion Matrix: is an evaluation metric to quantify the performance of a classification model. It shows the relationship between actual and predicted classes that can be used to calculate performance metrics such as precision, recall, and accuracy. The confusion matrix consists of four components as shown below:

	Positive (Actual)	Negative (Actual)
Positive (prediction)	True positive (TP)	False Positive (FP)
Negative (prediction)	False Negative (FN)	True Negative (TN

Mean Average Precision (mAP): calculated as the average of Average Precision (AP) values across all classes. It measures the accuracy of the model in detecting and locating the diagram elements in different categories. For example, it looks like the following**:**

$$mAP = \frac{1}{N}\sum_{i=1}^{N} \text{APi}$$

Character Error Rate (CER): is calculated as the ratio of the total number of character errors to the total number of characters in the reference text. It measures the accuracy of text recognition at the character level. It can be represented by the below.

$$\text{CER} = + \frac{\text{S} + \text{D} + \text{I}}{\text{N}}$$

Where **S** is the number of character substitutions, **D** is the number of character deletions, **I** is the number of character insertions, and **N** is the total number of characters in the reference (ground-truth) text.

Word Error Rate (WER) - calculates the ratio of the total number of word-level errors to total number of words in the reference text. It measures text recognition accuracy at the word level. It can be represented by the figures below

$$\mathrm{WER} = + \frac{\mathrm{S} + \mathrm{D} + \mathrm{I}}{\mathrm{N}}$$

where **S** is the number of word substitutions, **D** is the number of word deletions, **I** is the number of word insertions, and **N** is the total number of words in the reference (ground-truth) text.

BLEU (Bilingual Evaluation Understudy): This parameter measures the similarity of the generated description to the reference description using matching n-grams, and thus assesses the quality of the generated text by checking its similarity to the reference annotation as shown in the below

$$\mathrm{BLEU}=\mathrm{BP} \times \exp\left(\sum_{n=1}^{N} wn \log pn\right)$$

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): calculated as the similarity between the generated description and the reference description based on overlapping words or phrases. It evaluates the quality of generated text by measuring how much of the reference content is captured in the generated output. It can be represented by the below

$$\mathrm{ROUGE} = \frac{\text{Number of overlapping words}}{\text{Total words in Reference}}$$

CHAPTER 4: SYSTEM IMPLEMENTATION AND RESULTS

4.1. Introduction

This chapter describes the evaluation and implementation of our AI assisted system for converting STEM diagrams into accessible audio descriptions. This chapter describes how we constructed our idea in the last chapter, which then brought the design into practice.

The developed model is trained on a labeled STEM diagram dataset and trained on key diagram features such as symbols, text, arrows, and axes. It is assessed on the standard metrics of precision, recall, and mean Average Precision (mAP) to assess its effectiveness in detecting and classifying visual elements.

Beyond detection, this chapter also addresses the integration of different components into a single pipeline, which includes Optical Character Recognition (OCR) for text extraction, reasoning for semantic descriptions, and TTS for audio output. This complete solution enables us to turn diagrams into audio descriptions, thus making them accessible to visually impaired Students.

In the next few pages of this chapter, performance of the model is further described by way of training curves, confusion matrixes, and precision recall evaluations. Results from the qualitative testing on STEM diagrams are then presented. A web-based user interface is also provided through which the user can upload images and receive textual and audio outputs.

In sum, this chapter provides an overview of the implementation of the proposed system and demonstrates that it can bridge the gap between visual STEM content and learning by intelligent multimodal processing.

4.2. System Development Overview

This section illustrates the design and implementation of our AI-based system for converting STEM diagrams into accessible audio descriptions. The system is designed as an end-to-end multimodal pipeline that uses computer vision, optical character recognition (OCR), semantic interpretation, and text-to-speech (TTS) technology to convert visual information into meaningful audio output. Figure 4 illustrates the overall structure of our system. Image acquisition is followed by pre-processing and object detection using YOLOv12. Text regions identified are processed with

OCR to extract textual information. The outputs from detection and OCR are used in a layer of semantic interpretation to produce meaningful descriptions. Finally, the generated text is converted into audio via a text-to-speech module and delivered via a web-based user interface.

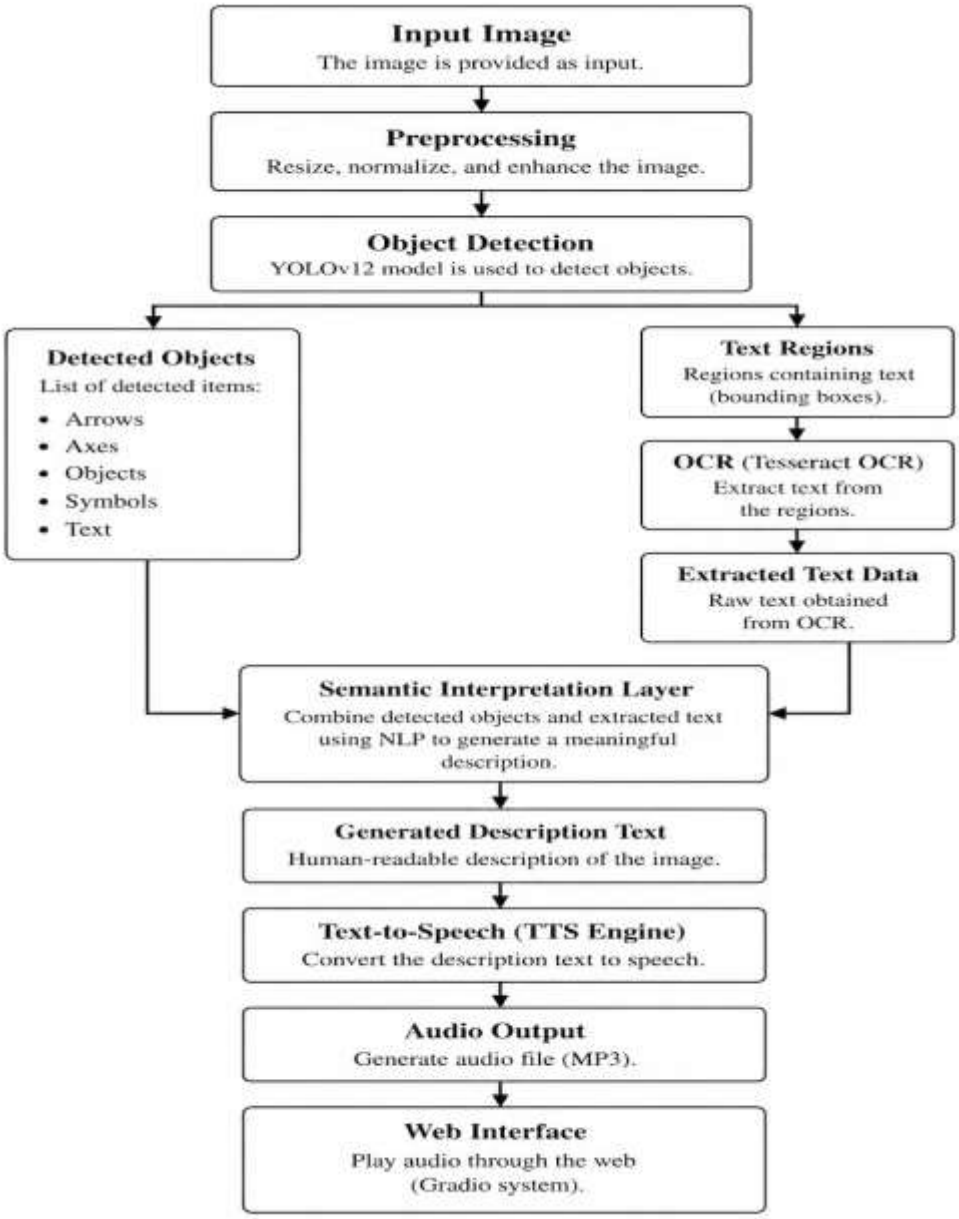


Figure 4.System Architecture of the Proposed JAIF Assistive Framework

Figure 4. shows the overall structure of the proposed JAIF assistive framework. The system includes image processing, object detection, OCR, semantic interpretation, and text-to-speech modules in one unified process to convert STEM diagrams into accessible audio descriptions.

The process begins with capturing input images such as digital textbooks, scanned diagrams, or user-uploaded educational content. These images are then resized, normalized, and adjusted to maximize visual appearance and consistency. Preparation of images is especially necessary due to the variety in resolution, lighting, and noise that is found in real world STEM materials.

Then, after some preprocessing, the system applies an object detection model based on the YOLOv12 architecture to identify and localize certain components in the diagram. The model was trained to detect five classes: arrows, axes, objects, symbols, and text. The detection process generated bounding boxes along with class labels and confidence scores, allowing a more thorough understanding of the diagram's visual elements.

Once a text segment has been identified, the OCR module extracts text. The OCR process converts image-based text into machine-readable text with Tesseract OCR. The extracted text is cleaned and normalized to reduce noise and make it easier to read, particularly in text with low resolution or stylized fonts.

In this stage, the results from object detection and OCR are combined to create a semantic description of the diagram. This is not the case when the detection system counts objects but uses context to identify the instructional nature of the diagram and its likely purpose such as identifying electrical circuits, graphs, or logical diagrams.

Finally, the description generated by the text is processed into speech using a text-to-speech engine and given feedback in the audio format, making it accessible to visually impaired Students. The sound output is clear and structured, reflecting the meaning of the diagram, rather than simply displaying data.

It is possible to browse through all of the pipeline information using a web interface developed in Gradio. The interface allows upload images, highlighted components, read text, and listened to audio explanations. The system was designed to be utilized as a tool with assistive technology applications.

Overall, the system is designed to be a cohesive and extensible platform that bridges the gap between STEM content displayed in images and engaging learning via deep learning and multimodal processing.

4.3. Dataset Preparation and Analysis

This section presents the preparation, structure, and analysis of the dataset used for training and evaluating the proposed object detection model. The dataset is comprised of annotated STEM diagram images representing common educational visual content (graphs, circuit diagrams, symbols)

We divided the dataset into training (70%), validation (20%), and testing (10%) subsets, with 140 images used for training, 40 images used for validation, and 20 images used for testing. Although 200 images may seem like a large data set for deep learning, the dataset was larger through transfer learning and data augmentation methods such as rotation, flipping, and scaling. These methods provided more diversity of data, improved model generalization, and reduced the need for a very large dataset while guaranteeing reliable training, validation, and testing performances.

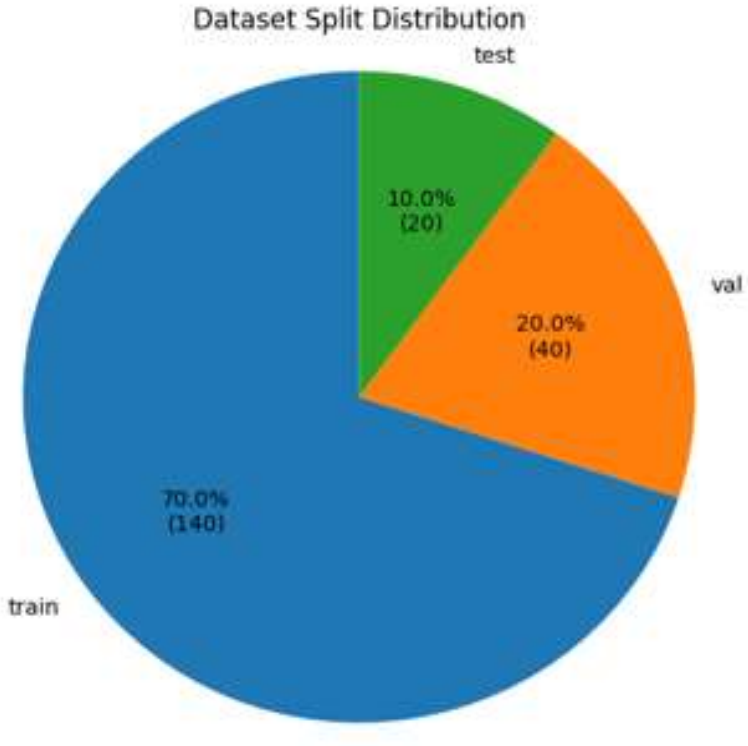


Figure 5.Dataset Split Distribution

Figure 5 plots the split of the data into training, validation, and testing sets. 70% of the data was used for training, 20% for validation, and 10% for testing to ensure proper model learning, parameter tuning, and independent evaluation.

Each image in the dataset is through a semi-automatic annotation process with bounding boxes for the five classes: arrow, axis, object, symbol, and text. These classes reflect the visual and semantic elements in STEM diagrams. The annotation ensures that structural components (such as symbols and objects) and informative components (such as text and directional signals) are represented correctly.

In contrast, the distribution of the classes suggests that the classes text and symbol have the most instances (318 and 296 annotations), followed by arrow and axis which are moderately represented, and object class with only 66 annotations.

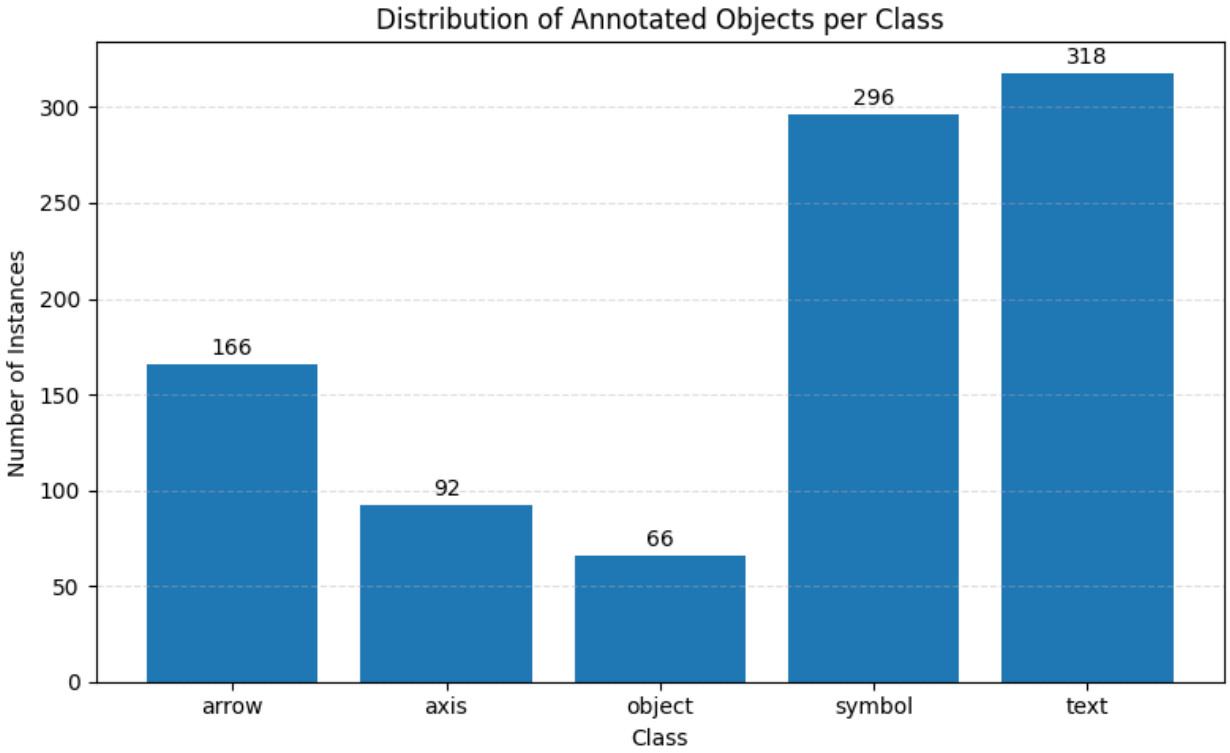


Figure 6.Distribution of Annotated Objects per Class

Figure 6 shows the number of annotated objects for each of the STEM classes in the STEM dataset. The distribution of arrows, symbols, text areas, and graphic elements used in model training is shown in the figure.

This is due to the natural composition of STEM diagrams in which textual labels and symbols are more frequent than objects by themselves. It also introduces challenges for model learning, particularly in classes that are not highly represented such as objects, which may lead to less detection.

To maintain consistency and improve model performance, all images are resized and normalized during preprocessing. The annotations are stored in YOLO format for fast training with the detection architecture chosen. The data quality is checked for label correctness, class mapping, and file integrity across all subsets of the image set.

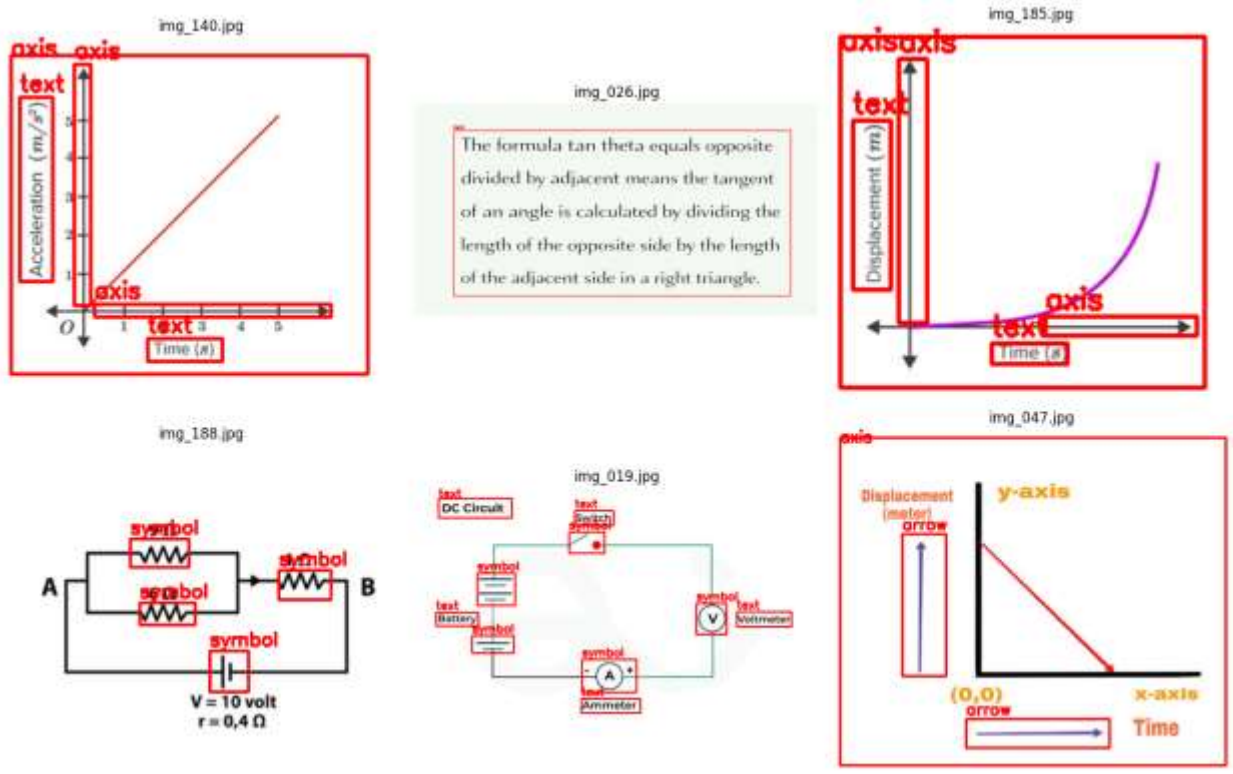


Figure 7.Sample Annotated Images.

Figure 7 shows sample of annotated STEM images used in supervised learning. The annotations indicate labeled items such as arrows, symbols, or text.

Overall, the dataset provides a broad representation of STEM diagram structures and can be used to develop a model that can detect both visual and textual components. The differences between

the classes represented and the size of the data set was be accounted for in the model analysis and discussion in the following sections.

4.4. Model Training and Performance Evaluation

This section describes the training and evaluation of the object detection model for detecting components in STEM diagrams. The model is implemented using YOLOv12 and trained on the prepared dataset to detect five classes of components: arrow, axis, object, symbol, and text. Evaluation is made using both quantitative measures and visual measures.

4.4.1. Training Behavior

We evaluate the training process of YOLOv12 model with loss functions and performance metrics across epochs. The model is trained on a STEM diagram to learn object localization and classification for five classes.

Table 3.Training Configuration

Parameter	***Value***
Model	YOLOv12
Epochs	50
Image Size	640 × 640
Batch Size	16
Optimizer	SGD (default Ultralytics)
Learning Rate	Automatic scheduling
Hardware	Colab

Table 4.Training Loss Behavior

Loss Type	***Observation***
Box Loss	Decreases steadily, indicating improved localization accuracy
Classification Loss	Rapid reduction, showing effective class learning
DFL Loss	Smooth decline, improving bounding box refinement

The following figure depicts the training and validation curves. It depicts the evolution of loss functions and evaluation metrics over training epochs. The training results show a steady decrease in all loss components. The box loss decreases from about 2.2 to below 1.0, which indicates better object localization. The class loss also decreases, indicating more correct classification of the five classes. The distribution focal loss also decreases slowly, indicating better box bounding box refinement.

The validation losses exhibit the same pattern with very minor variations in the early epochs, normal due to the complexity of data and class imbalance. The validation losses converge to the training losses in later periods without significantly diverging, indicating that there was no significant overfitting.

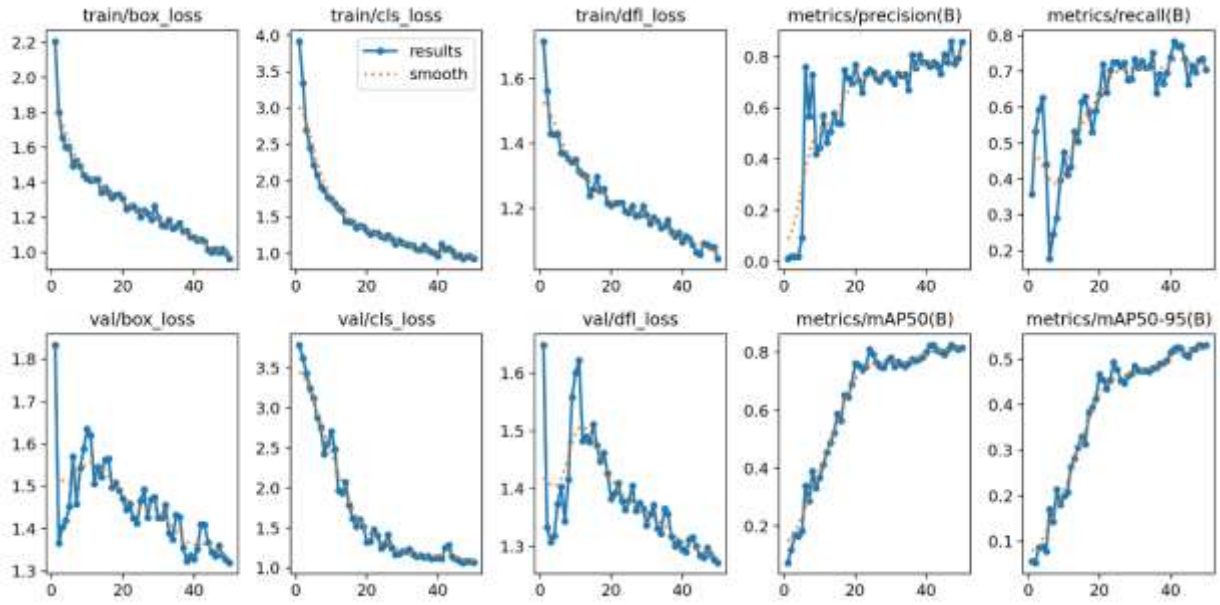


Figure 8.Training and Validation Curves

Figure 8 illustrates the training and validation performance curves of the proposed model during training iterations. Not only does the model reduce losses, but the results improve with each epoch of training as well. The accuracy curve increases rapidly in the early epochs and eventually stabilizes at 0.75, while the recall curve slowly climbs and stabilizes around 0.73, meaning the model is balancing false positives and false negatives.

The mAP@50 curve increases linearly and converges at approximately 0.81, indicating that mAP@50–95 detects well at the standard IoU level. The mAP@50–95 curve also increases linearly to approximately 0.53, which implies that it detects well under more stringent localization conditions.

As expected, the overall train-out response is smooth without divergence or overfitting. The linear transitions of the train- and validation-rates suggest that the model generalizes well to unseen data. But, slight fluctuations in the validation loss and recall may be due to a class imbalance, such as higher representation of some classes, such as objects.

Overall, the training phase of the model demonstrated that it successfully learned meaningful representations of the features from the data set and had stable convergence and balanced detection across several performance measures.

4.4.2. Detection Performance Metrics

Our models are trained with a YOLOv12 model. We test for object detection precision, recall, mean Average Precision at IoU threshold 0.5 (mAP@50), and mean Average Precision across multiple IoU thresholds (mAP@50–95).

Table 5.Overall Detection Performance

Metric	***Value***
Precision	0.768
Recall	0.735
mAP@50	0.815
mAP@50–95	0.530

Overall precision is 0.768 and recall is 0.735, which indicate a balance between recognizing the object from its context and preventing false positives. The mAP@50 score of 0.815 indicates a high accuracy of detection in normal overlap scenarios, while the mAP@50–95 score of 0.530 indicates moderate robustness in localization situations.

Plus, these results indicate that the model is effective in identifying key elements of STEM diagrams at common thresholds of evaluation. But its performance falls precipitously at higher thresholds of

IoU, suggesting that though the object is found reliably, it remains somewhat difficult to identify the bounding boxes within the STEM diagram.

To further validate these quantitative predictions, we examine our qualitative prediction outputs to see if our model is able to correctly predict symbols, text areas, and objects in various diagrams.

4.4.3. Class-wise Performance Analysis

Class-wise results reveal that detection performance varies across categories due to differences in visual structure, scale, and representation complexity.

Table 6.Class-wise Detection Performance

Class	***Precision***	***Recall***	***mAP@50***	***mAP@50–95***
Arrow	0.716	0.600	0.756	0.347
Axis	0.768	0.765	0.835	0.594
Object	0.873	0.690	0.787	0.633
Symbol	0.765	0.870	0.917	0.606
Text	0.716	0.752	0.780	0.473

The most highly precise and recall values are obtained for the symbol class, which indicates that the model learned consistent geometry such as resistors, circuit elements, or standard diagram symbols, which have well defined shapes and boundaries and are easier to detect. The axis and object classes perform fairly well, with relatively consistent visual features that vary only in orientation and placement. The text class performs moderately poorly due to the variable size, style, position, and embedding of text in graphics. Unlike symbols, text regions do not have consistent structural structure so are harder to detect and to localize accurately. The arrow class is relatively low performance because of its small size, thin structure, and significant variability in direction and location, reducing detection confidence and confusion about its localization.

4.4.4. Confusion Matrix Analysis

The confusion matrix quantifies the classification behavior of the model by comparing predicted labels with ground truth annotations. The raw and normalized confusion matrices are used to analyze the detection accuracy and misclassification across classes.

The confusion matrix summarizes object-level predictions generated during evaluation. Although only 20 images were used for testing, each image contained multiple annotated STEM images components. Consequently, the evaluation involved approximately 184 object instances across the five target classes. The higher values in the confusion matrix are due to multiple object instances being detected and classified in the test images. The image split was performed at image level, not object level. Since diagrams contain different numbers of annotated objects, the object distribution across train, validation, and test sets is not exactly proportional.

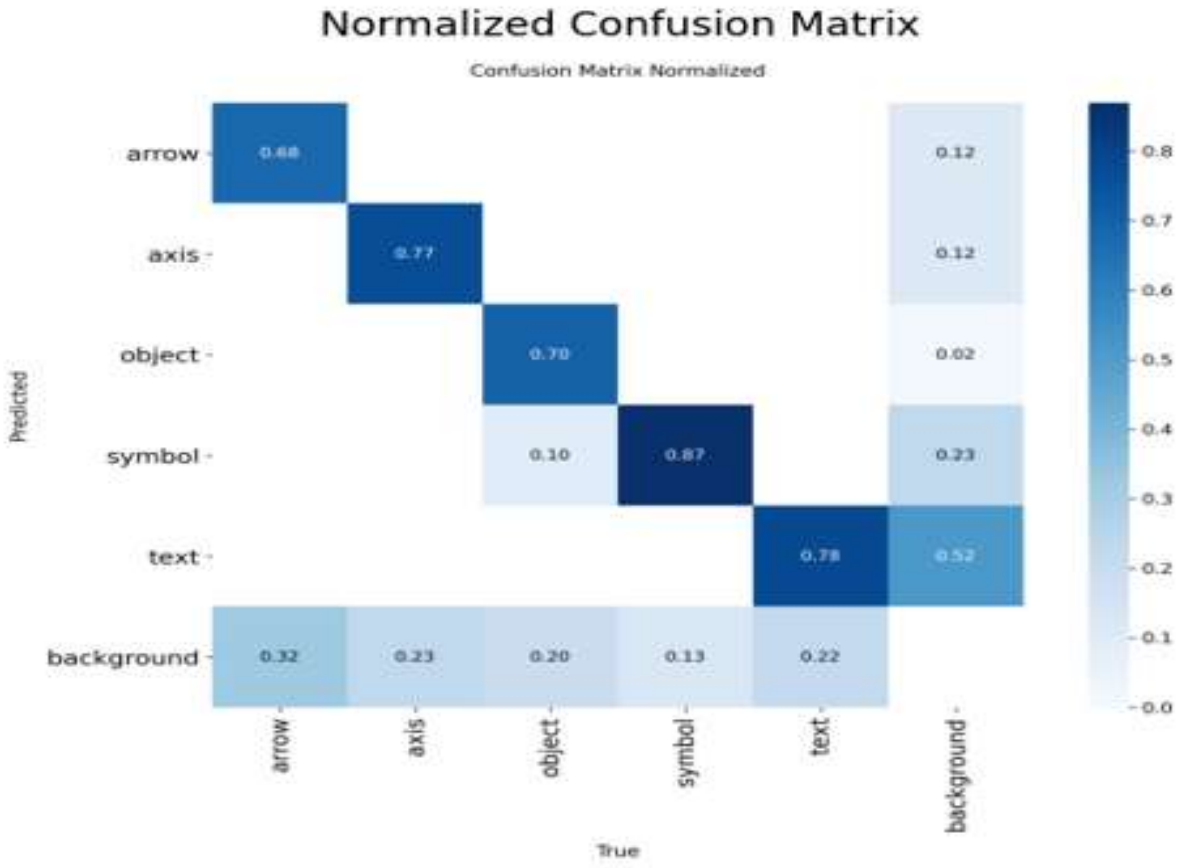


Figure 9.Normalized Confusion Matrix

Figure 9 illustrates the confusion matrix of the object detection model. This matrix shows the relationship between actual and predicted classes. This normalized confusion matrix reveals strong diagonal values for most classes, which indicates that the model correctly classes most instances. The symbol class has the highest classification consistency (0.87), which is consistent with

previous scores. The text class has a relatively high correct classification rate (0.78), indicating that even when text looks different from the sample text, the model correctly classifies the text.

Although there are some notable misclassification patterns, some confusion exists between the symbol and object classes, and visually similar elements may sometimes be misclassified. This is evident by the off-diagonal values, which reflect incorrect predictions about these classes. And finally, some text cases are partly or completely missed due to low contrast, small size, or being hidden in complex diagram structure.

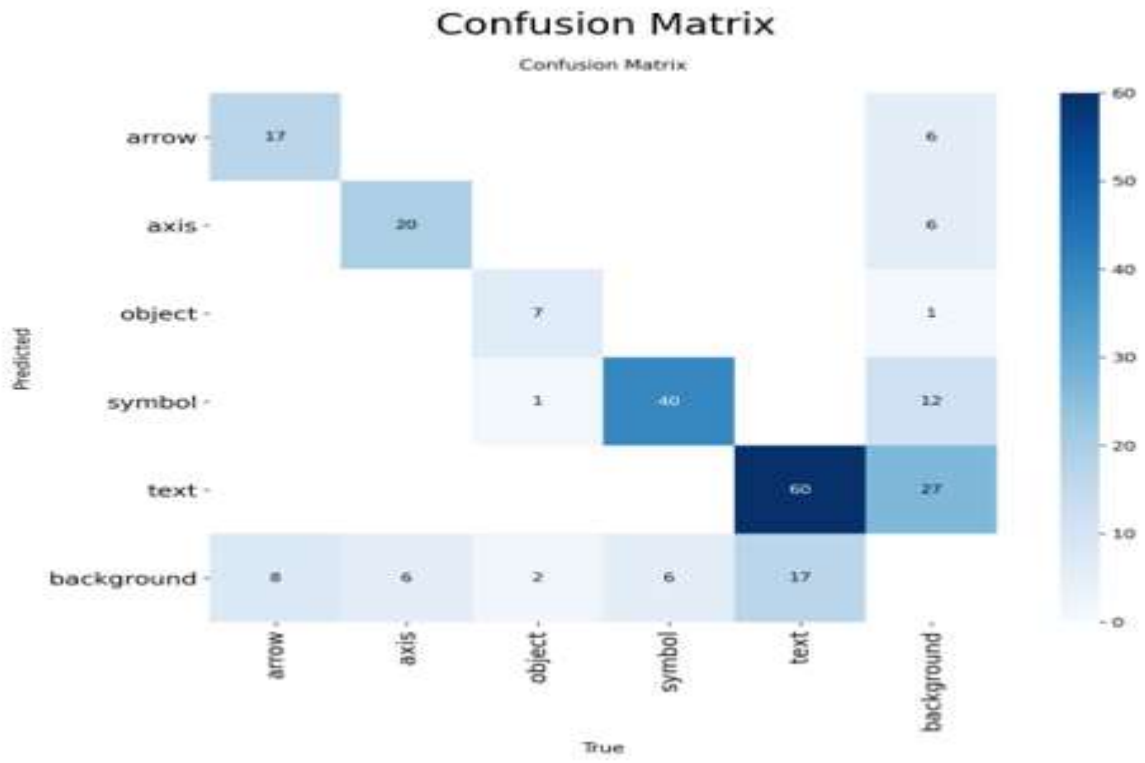


Figure 10.Confusion Matrix

Figure 10 shows the normalized confusion matrix used to evaluate classification accuracy in different object categories. The raw confusion matrix also shows these patterns, but in absolute numbers. For example, while 60 instances of text were correctly identified, several were misclassified as background (17 instances). This reinforces the difficulty of distinguishing text from the noise.

Background misclassification across several classes indicates that some objects are not confidently detected and produce false negatives. This is common in dense, visually complex diagrams where components overlap and annotation quality differs.

Overall, the confusion matrix results show that the model is very reliable in detecting properly structured and defined elements, but it performs poorly in classes where there is small size, visual variability, or context uncertainty, such as arrows and text.

4.4.5. Precision–Recall and Confidence Analysis

To further evaluate the detection behavior of the model at various decision thresholds, precision-recall and confidence-based performance curves are also computed. These curves provide a deeper insight into how the model balances detection accuracy and completeness in different operating conditions.

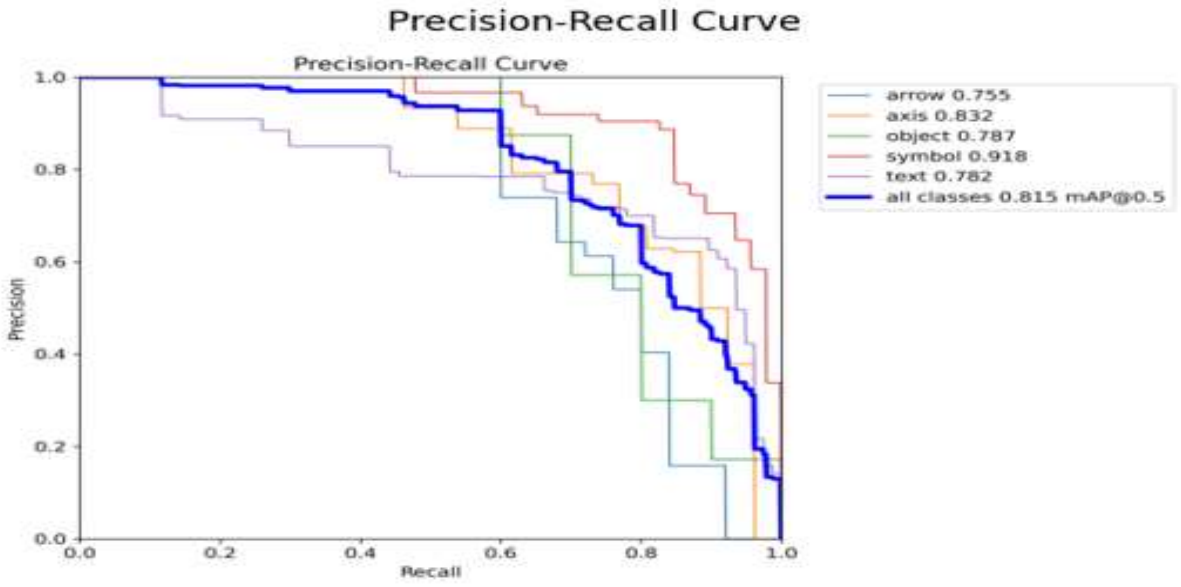


Figure 11.Precision–Recall Curve showing the relationship between precision and recall across different classes.

Figure 11 plots the Precision-Recall curve of the proposed model for the balance between detection precision and recall. The precision-recall curve indicates that the model is relatively accurate for a broad range of recall values. This suggests that the model is able to properly identify relevant objects while preventing false detections. Overall, we show a performance similar to mAP@50 of

about 0.815 with a mean precision close to that reported earlier. But, as recall gets larger, we find that the precision decreases, particularly in more complex classes such as arrow and text. These results suggest that an increase in detection coverage introduces more false positives, especially for small or visually variable items.

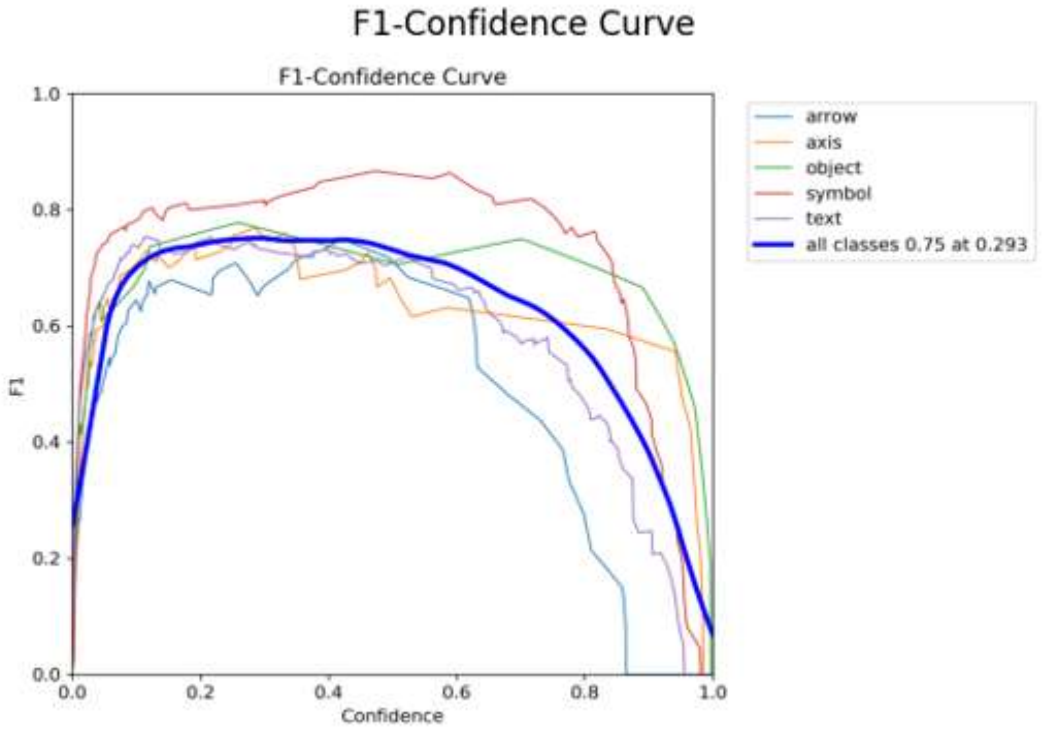


Figure 12.Confidence Curve illustrating the optimal confidence threshold for balanced detection performance.

Figure 12 shows the confidence curve to determine the optimal confidence threshold for object detection. The F1–confidence curve further illustrates the trade-off between precision and recall. It shows that the optimal balance is achieved at a confidence level of 0.29. This is the best confidence threshold for the model, i.e., the limit of accuracy of detections and object coverage. Lower confidence levels improve recall more, but less accurately, while higher confidence levels improve precision but not accuracy.

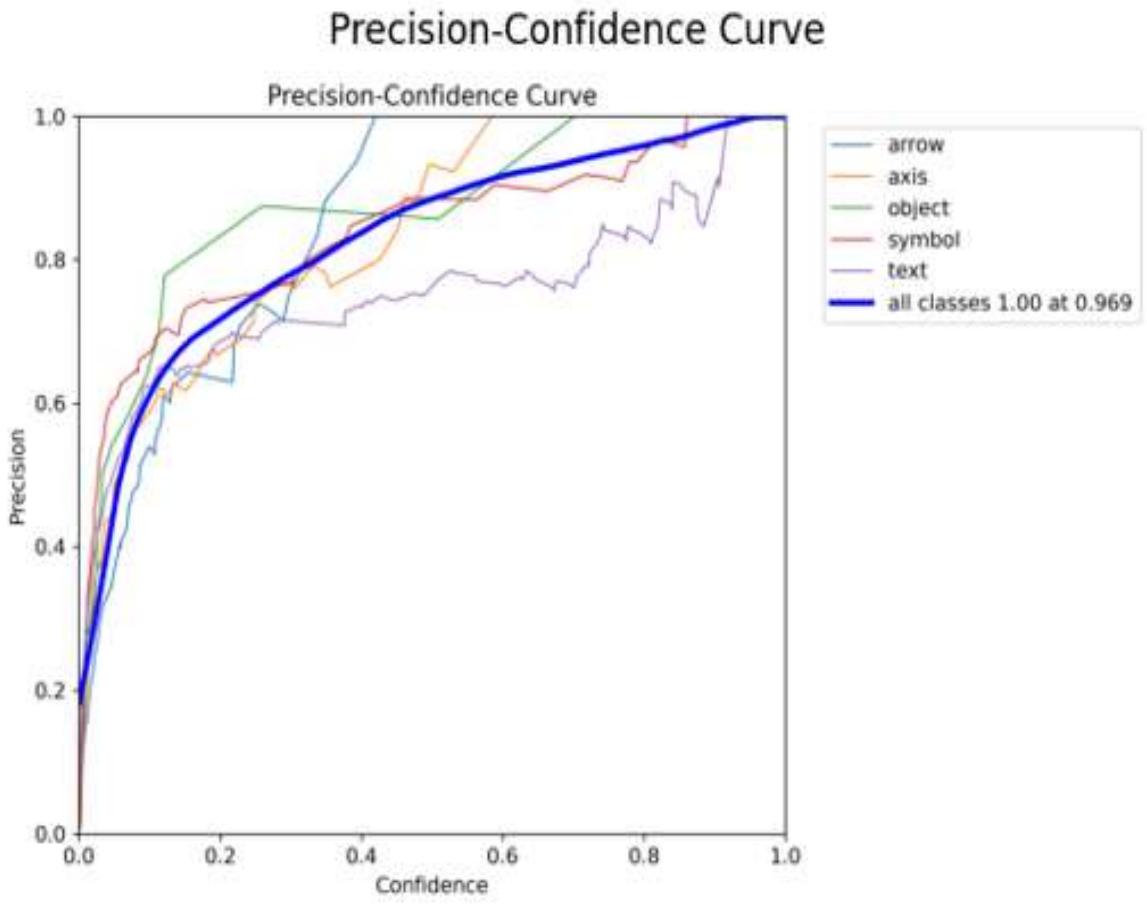


Figure 13.Precision–Confidence Curve demonstrating the increase in precision with higher confidence thresholds.

Figure 13 shows the relationship between prediction confidence and precision performance. The precision–confidence and recall confidence curves illustrate this relationship even more strongly. As the confidence threshold is increased, precision increases steadily, approaching near-perfect values at high thresholds. In contrast, recall decreases significantly, indicating that less objects are detected as the model becomes more conservative. This effect is particularly important in applications where the choice of confidence threshold directly influences system behavior.

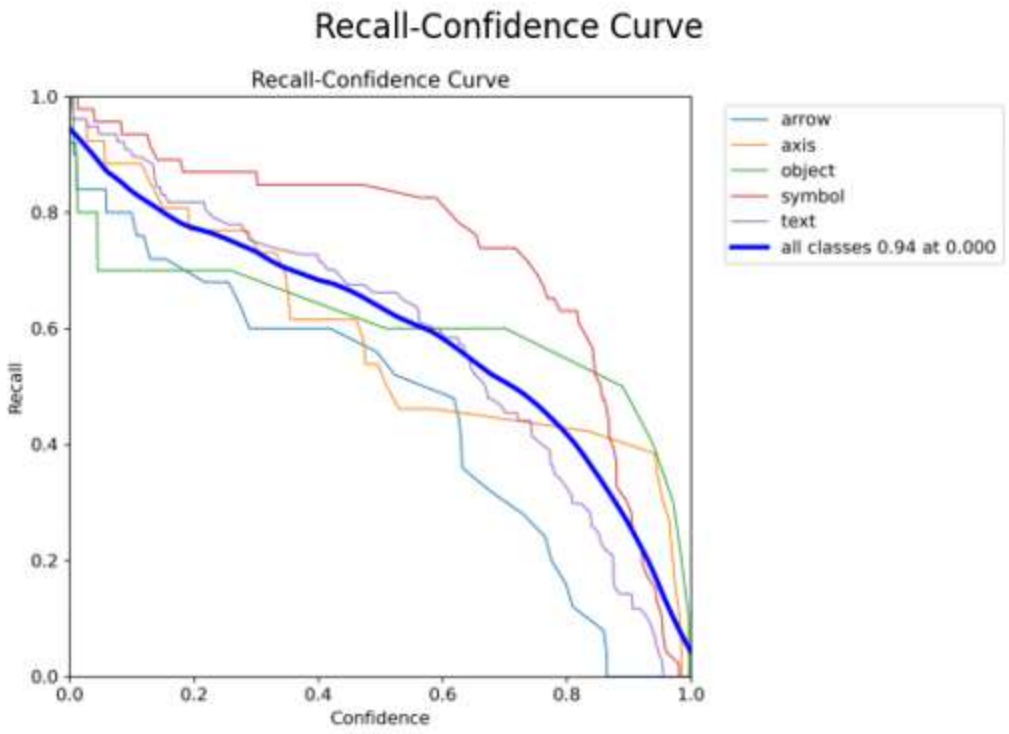


Figure 14.Recall–Confidence Curve showing the decrease in recall as the confidence threshold increases.

Figure 14 depicts the Recall–Confidence relationship in the proposed model. Notably, the recall curve declines more sharply for classes like arrow and text, which indicates that these categories are more susceptible to confidence threshold adjustments because they are smaller, have lower visual contrast, and are more variable, which decrease detection confidence. Although more structured classes like symbols perform relatively well over a wider confidence range, this is not true for larger classes like fonts.

Overall, these results indicate that the model behaves predictably at low confidence levels. The best value is in a moderately low confidence level to prevent misses of key parts of the diagram, even if tiny false positives are present.

4.4.6. OCR Performance Metrics

The OCR module was evaluated using Character Error Rate (CER) and Word Error Rate (WER) to assess the accuracy of the text extracted from STEM diagrams. Samples were prepared with

scientific labels, numerical values, instructional text, and electric notations. Ground-truth annotations were manually generated and compared to the OCR outputs generated by JAIF.

Table 7.Sample OCR Evaluation Results

Sample	***Ground Truth Text***	***OCR Output***
Sample 1	R = 3 Ohms, E = 12 Volts, A simple circuit diagram, 4 Amperes	R3 Ohms, E 12 Volts, A simple circuit diagram, 4 Amperes
Sample 2	5.02, 9.0 V, 3.00	5.02, 9.0 V, 3.00
Sample 3	opposite, adjacent	opposite, adjacent
Sample 4	10 N to the left, 10 N to the right, 60 N to the left, 50 N to the right. What is the net force on the box shown below?	aio n to the left 10 N to the right 60 N to the left 50 N to the right, what is the net force on the box shown below, what is the net force on the box

Table 7 presents sample OCR results by comparing the ground-truth text with the corresponding text extracted by the OCR module. The table provides a qualitative assessment of OCR performance across different STEM diagram samples and illustrates the system's ability to accurately recognize labels, annotations, and textual elements embedded within visual educational content.

The results indicate that the OCR module can recognize simple and clearly structured text. Samples 2 and 3 were recognized perfectly with CER and WER values of 0.000. Therefore, the OCR module can extract short and distinct text from STEM diagrams.

Sample 1 had a low CER value of 0.042 because there was only one or two character-level errors in recognition. In contrast, WER is 0.286 because small characters were missed or formatting differences affected complete word matching. This shows how sensitive WER is to partial word-level errors.

Sample 4 had the highest CER of 0.356 and WER of 0.433. It had lengthy instructions with dense layout and many directional labels, making recognition more difficult. The OCR output often included merged words and character distortions, which further aggravated both character- and

word-level errors. Despite these challenges, the text still retained much of the semantic information needed for the description.

Overall, the results of the OCR test show that the proposed system can extract text reliably from structured STEM diagrams, but fails when text is noisy or dense.

Table 8.OCR evaluation results

Sample	***CER***	***WER***
Sample 1	0.042	0.286
Sample 2	0.000	0.000
Sample 3	0.000	0.000
Sample 4	0.356	0.433

Table 8 presents the quantitative evaluation of OCR performance using Character Error Rate (CER) and Word Error Rate (WER). These metrics measure the accuracy of text extraction by comparing OCR-generated text against the ground-truth text. Lower CER and WER values indicate higher recognition accuracy and better overall OCR performance within the JAIF framework.

Table 9.Summarized OCR evaluation metrics

Metric	***Result***
Average CER	0.099
Average WER	0.180

OCR extracted information perfectly from schematics representing STEM subjects with simple labels and numbers. These examples of simple diagrams included an electric circuit and a coordinate system. OCR did not perform as well on complex instructional diagrams with numerous annotations, dense layouts, small font sizes, or noisy text.

As we read the evaluation sheets, we noted some problems with separating words into smaller fragments, duplicate phrases, sentence segments cut off, and incorrect character segmentation. These were particularly noticeable in force diagrams and instructional-question figures where multiple sections of text overlapped with graphic parts and arrows.

Despite these limitations, the OCR results were sufficiently accurate to support the interpretation and description of meanings as proposed in the proposed JAIF. The textual information extracted assisted in understanding context and a more accessible narration for visually impaired participants. Figure 15 depicts the OCR extraction and generated audio outputs from STEM diagrams.

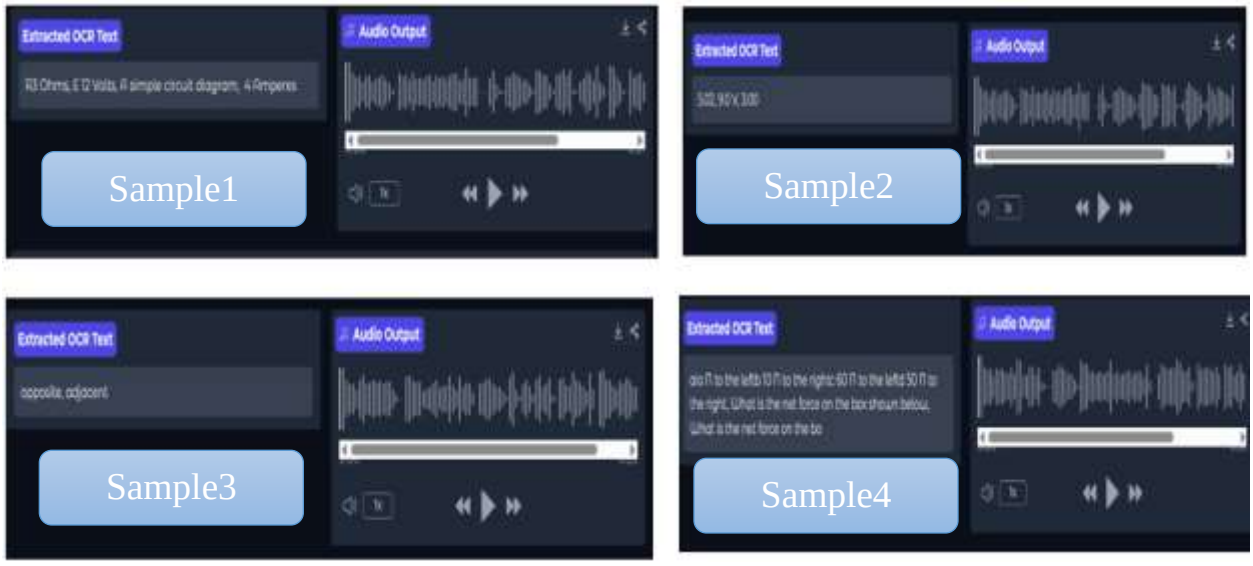


Figure 15.OCR Extraction and Assistive Audio Outputs

4.4.7. NLP Performance Metrics

The natural language processing (NLP) module was assessed using BLEU and ROUGE metrics, which measure the semantic similarity of generated assistive descriptions to manually drafted reference descriptions. Evaluations focused on how well JAIF converted STEM components and OCR information extracted into more easily recognizable sounds for visually impaired Students.

Table 10.Sample NLP Evaluation Results

Sample	***Reference Description***	***Generated Description***
Sample 1	The diagram shows an electrical circuit containing voltage, resistance, and current values connected in a simple circuit.	This image appears to be an electrical circuit diagram. It contains components connected together to represent electrical behavior such as voltage or resistance.

Sample 2	The diagram represents a graph with labeled axes and plotted relationships between variables.	This image appears to be a graph or coordinate-based STEM diagram. It shows a relationship between variables using axes and plotted information.
Sample 3	The diagram contains directional arrows and instructional force-related labels.	This image appears to be a general STEM instructional diagram with labeled components and directional arrows.
Sample 4	The figure contains resistor symbols and labeled electrical components.	Scientific symbols represent components or operations. The readable text includes resistor and electrical labels.

Table 10 summarizes the NLP evaluation metrics. In terms of semantic similarity, NLP module successfully generated structured descriptions of STEM diagrams, the BLEU score of 0.116 indicates limited lexical similarity to the reference descriptions. Therefore, the results should be interpreted cautiously, as BLEU may not adequately capture semantic equivalence in assistive descriptions. Future studies should incorporate ROUGE-L and human evaluation by educators and visually impaired users to provide a more comprehensive assessment of description quality.

Table 11.Summarized NLP evaluation result

Metric	***Result***
BLEU Score	0.116
ROUGE-1	0.505
ROUGE-2	0.265
ROUGE-L	0.440

In contrast, the generated and reference descriptions had some semantic overlap (within a few words) in the ROUGE based assessment, suggesting that the system managed to keep important instructional content and context information from STEM diagrams. Despite small differences in sentence structure, the generated descriptions accurately referred to categories of diagrams, detected elements, meanings for instruction, and scientific context.

The NLP module performed well on diagrams representing structured STEM information such as electrical circuits or coordinate-based educational figures where components were connected

semantically. The outputs described electrical behaviors, graph structure, axes, plotted relationships, and symbols from science in a way that was easy to hear.

In diagrams with noisy OCR outputs, incomplete detections, or dense instructional annotations, some of the described concepts were repeated, simplified, incomplete, or confused through OCR extraction and overlap detected objects.

Although there were some difficulties with understanding the generated descriptions, they were understandable and informative, and the framework presented here could be adapted to help the learning of accessible STEM. Using object detection, OCR extraction, context knowledge, and narration, the JAIF framework made it possible for the visually impaired Students to listen to the explanations of the visual content. Figure 16 shows accessible descriptions that were generated by the NLP module for visually impaired Students.

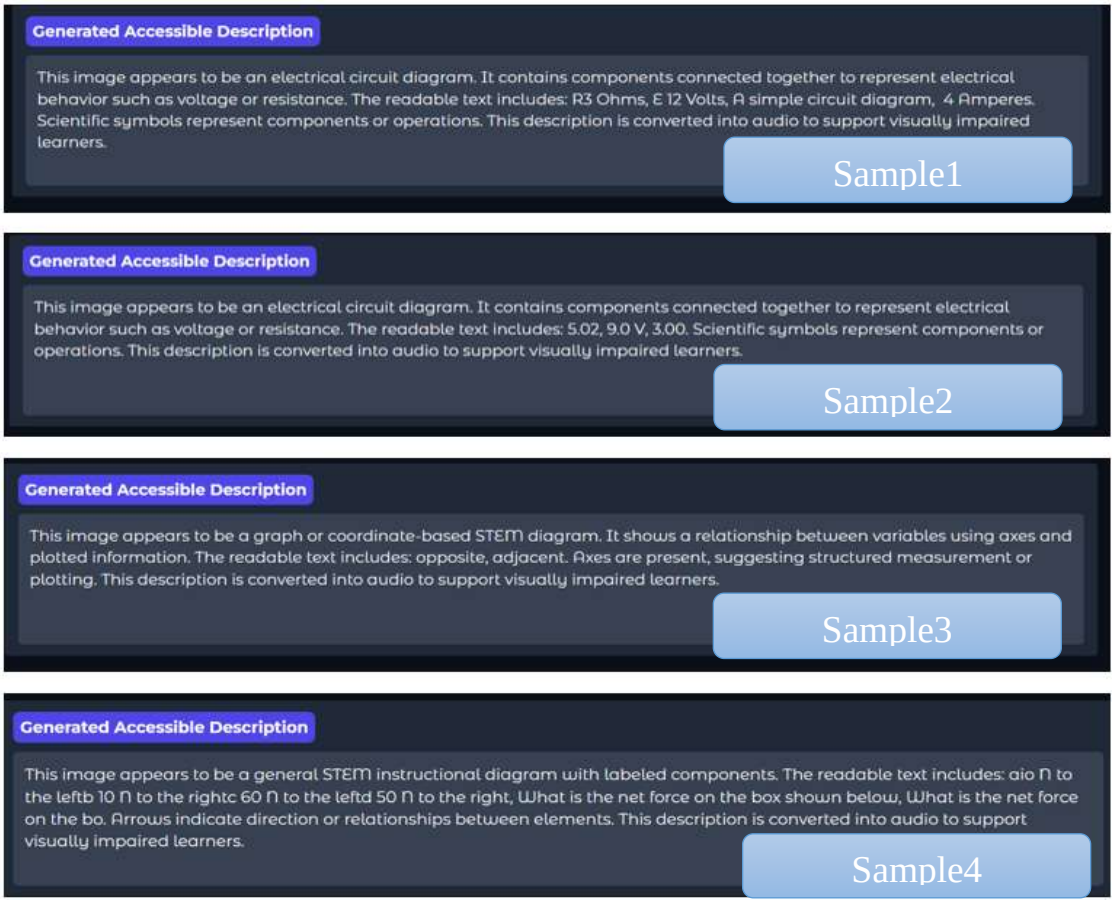


Figure 16.Generated Accessible Descriptions

4.5. Qualitative Results and Visual Evaluation

In addition to quantitative analysis, we performed qualitative analysis on representative prediction outputs of the trained model on unseen STEM diagram images. These visual results provide an insight into how the model performs in the real world, showing its strengths and weaknesses.

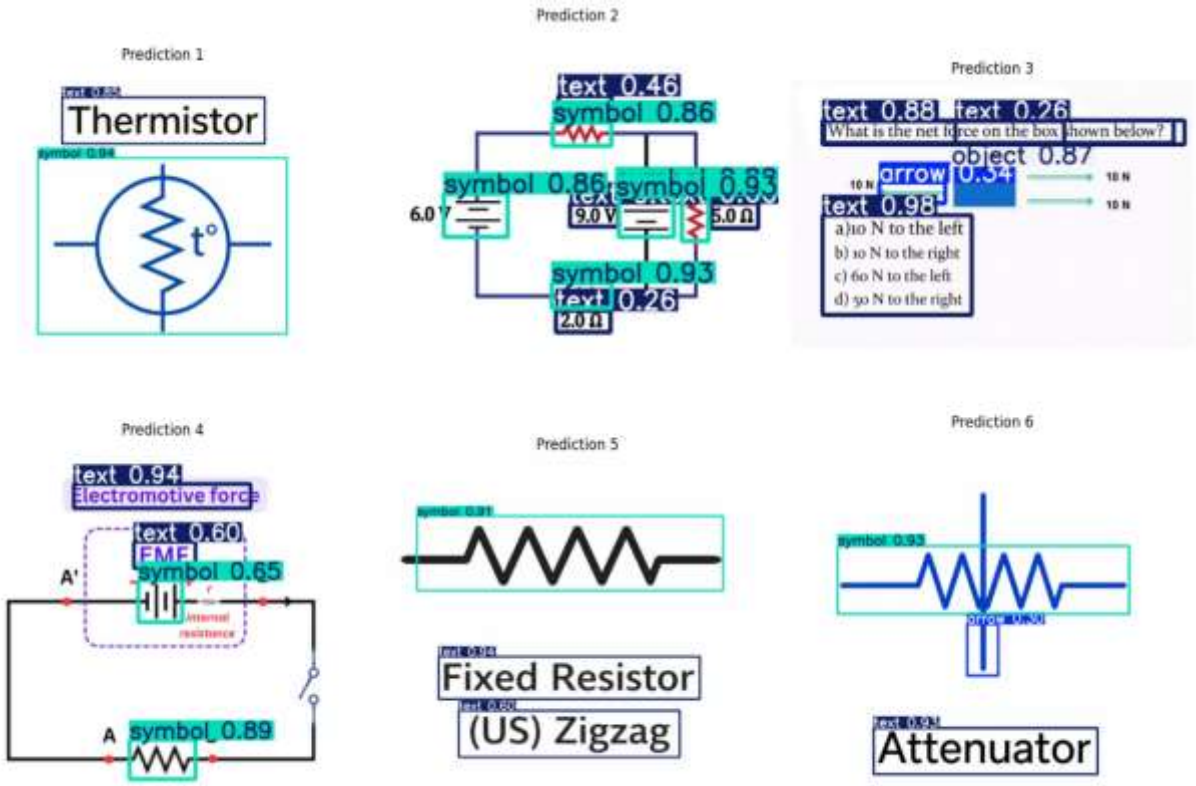


Figure 17.Sample qualitative detection results showing model predictions on diverse STEM diagrams, including symbols, text regions, arrows, and objects.

Figure 17 shows qualitative detection results of the proposed model on different STEM diagrams. Several examples indicate that circuit elements, such as resistors and electrical components can be identified in the model with high accuracy, as demonstrated in the quantitative analysis for the symbol class. This indicates that the model knows the geometric pattern associated with structured diagram elements.

Text can be detected in any type of diagram, including labels such as “Thermistor,” “Electromotive force,” and “Fixed Resistor”. Unfortunately, border boxes can be a bit finicky and may overlap occasionally, especially when a diagram contains lots or layers. This behavior is expected given

the moderate detection performance of the text class, where font style, size, and orientation can be tricky to detect.

Detection of arrows and direction indicators is present, but less consistently. In some cases, arrows are correctly identified, in others, they are detected with less confidence or partially localized. This is expected given the relatively lower performance metrics for the arrow class because of their smaller size, thin structure, and orientations in diagrams.

More complex diagrams with multiple elements pose more problems. Sometimes overlapping detections and closely located elements result in ambiguous bounding boxes, especially between symbol and object categories. These are typical patterns of confusion in the confusion matrix where visually similar features lead to misclassification.

Despite these limitations, the qualitative results show that the model is well-suited to describe STEM diagrams with circuit diagrams, components labeled, and instructional images. The model correctly identifies visual elements that are necessary for text extraction and interpretation.
Assistive aspects of the results, these are significant. Localization of symbols and text sections is crucial in extracting relevant information for descriptions. Even though small detection errors may occur for smaller or less distinctive elements, the model allows sufficient structural understanding of diagrams to aid the next steps of the pipeline, i.e., OCR and audio transcription.

4.6. System Integration and Assistive Output

In this section, we present the integration of the components into a unified assistive system for turning STEM diagram images into accessible audio descriptions using object detection, optical character recognition (OCR), semantic interpretation, and text to speech in a single interface.

Figure 5 shows how the system starts with user input of a STEM diagram that is uploaded to the system through the website. The uploaded image is processed by the trained object detection model that identifies and places key elements of the STEM diagram such as symbols, text area, arrows, and objects on the original image, providing immediate feedback on the performance of the model and for the user to check the results. Figure 18 illustrates the user interface of the developed JAIF assistive system.

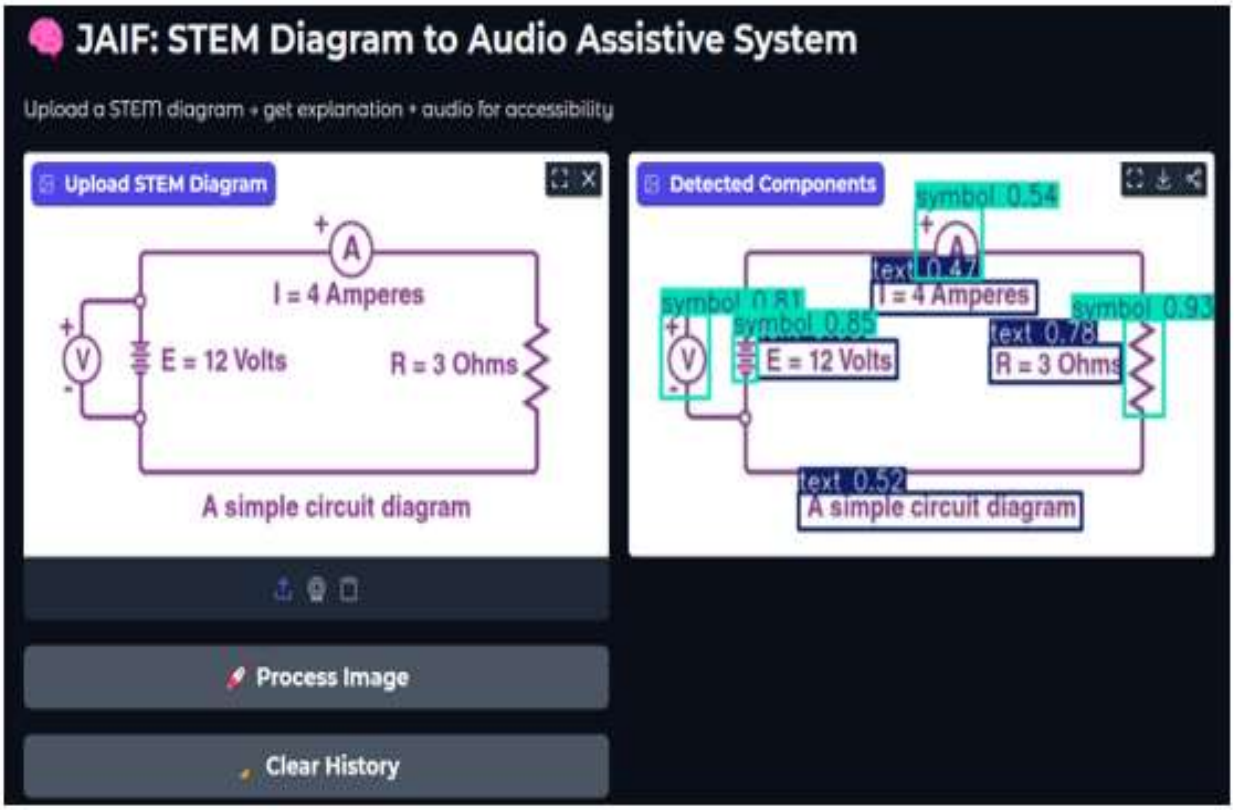


Figure 18.System interface showing image upload and detected diagram components.

As the detected areas are located in the diagram, text is extracted from them through OCR, such as the number, label, and annotations they contain. As shown in the system output, the fields were correctly identified: "R 3 Ohms," "E 12 Volts," and "I 4 Amperes." This is critical because this step bridges the gap between the visual content and the underlying meaning. Figure 19 shows the final output of the proposed framework in terms of assistive information.

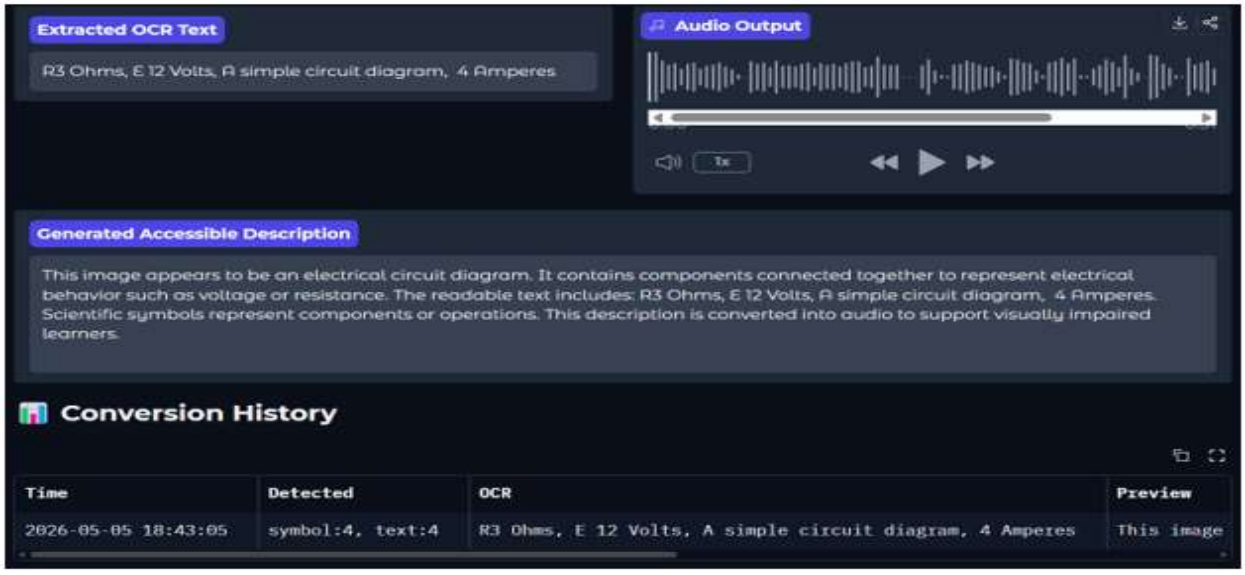


Figure 19.OCR extraction, generated accessible description, and audio output from detected diagram elements.

The visually and textually extracted information is then passed to the semantic interpretation layer, where a description of the diagram is generated. This is not just a list of detected elements, but also an explanation that places components in a meaningful meaning. For example, the system identifies the diagram as an electrical circuit and describes its components and values in natural language. This represents the transition from low-level perception to higher-level representation, which is useful for users.

The final step is to convert the text description into an audio signal by using a text-to-speech component. The device emits an audio signal that allows visually impaired Students to read the diagram content through the auditory medium. This is the primary goal of the proposed system, which is to enable visually impaired Students to access multimodal information via a visual diagram.

This process takes place in real-time, but the system keeps a history of previous conversions (Figure 19). This feature facilitates user interfaces by allowing a user to review past interactions with the system, such as detected components, extracted text, and generated descriptions. This also indicates that the system operates as a complete application instead of a model that allows for continual user interaction.

Overall, the integrated system demonstrated the feasibility of converting STEM diagrams into accessible audio descriptions using a structured processing process. Although the system still has some issues with detection precision and semantic depth, it provides a functional and scalable platform for assistive technologies that help accessibility in STEM education.

4.7. Discussion

This study focused on the design and evaluation of the JAWS–AI Integrated Assistive Technology Framework (JAIF) to improve the accessibility of STEM educational materials for visually impaired students. Results revealed that the proposed system correctly interpreted STEM diagrams and transformed visual educational content into structured auditory descriptions using Artificial Intelligence and assistive technologies. The integration of object detection, OCR, NLP, TTS, and JAWS demonstrated that the proposed AI modular structure was suitable for the real-time interpretation of STEM diagrams.

The proposed JAIF architecture successfully included object detection, Optical Character Recognition (OCR), Natural Language Processing (NLP), Text-to-Speech (TTS), and JAWS screen reader in a single assistive system. From image input to output of audio, the proposed architecture was able to comprehend STEM diagrams in real time.

YOLOv12 detected and classified diagram items such as arrows, symbols, axes, objects, and text. The YOLOv12 model detected and classified most diagram elements correctly, such as arrows, symbols, axes, objects, and text. The high accuracy, Precision, recall, and mAP values were highest for different classes of diagrams. In addition, the confusion matrix and precision–recall curves indicated that most diagram elements correctly identified with a few misclassifications. These results indicated that the object detection module could interpret structured visual educational images. OCR data indicate that the framework extracts text from STEM diagrams with high Character Error Rate (CER) and Word Error Rate (WER) values, indicating that it correctly recognized labels, annotations, and numerical values. The extraction of accurate text supported the understanding of the diagrams as a whole.

The NLP scores also revealed the sense, structure, and context-awareness of the generated descriptions. BLEU and ROUGE scores showed high similarities between the generated descriptions and the reference descriptions. Rather than creating labels for all parts of the diagram,

the model generated explanations that connected parts of the diagram. The description of the STEM visual elements was more accessible and understandable for visually impaired Students. Another example was the integration of the JAIF framework and JAWS screen reader. The generated text descriptions were converted into speech from Text-to-Speech and delivered through JAWS in real time. This indicated that the proposed framework moved screen readers beyond just text reading and toward intelligent interpreting of graphics.

Overall, the system performed well in terms of detection accuracy, text extraction accuracy, semantic description quality, and processing efficiency. Across all evaluation metrics such as precision, recall, mAP, CER, WER, BLEU, and ROUGE, it performed well in detecting, interpreting, and describing STEM diagram content.

Although the design had some limitations such as a small number of images, variations in image quality, and domain-specific attention, it was found that the approach could be used for inclusive learning. Overall, the study showed that object detection, OCR, NLP, and assistive technologies can improve accessibility to visual educational material for visually impaired students and inclusive STEM education.

CHAPTER 5: CONCLUSION AND RECOMMENDATION

5.1. Conclusion

This research study aimed to create and implement an AI-Integrated Assistive Technology Framework (JAIF) designed to make learning accessible for visually impaired students, particularly with STEM visual instructional materials. Present assistive technologies such as screen readers allow access to textual information, but they are not capable of interpreting diagrams, charts, graphs, or other visual aids used in science and technical education. This creates barriers for visually impaired students and affects their independence in inclusive learning environments.

Toward this end, the study designed and implemented the JAWS-AI Integrated Assistive Framework (JAIF), a combination of deep learning-based object detection, Optical Character Recognition (OCR), Natural Language Processing (NLP) and Text-to-Speech (TTS) in one assistive framework. The framework is intended to turn STEM diagrams into structured and context-aware auditory descriptions which could be translated to assistive screen readers.

The system consists of image preprocessing, object detection (YOLOv12), text extraction (OCR) and semantic interpretation (audio generation). It was trained on annotated STEM diagram sets and evaluated under controlled conditions. Results showed that our system could be used to detect diagram elements, extract text, create meaningful descriptions, and generate audio output.

Across the different evaluation measures, the proposed JAIF system achieved a precision of 0.768, recall of 0.735, mAP@50 of 0.815, mAP@50–95 of 0.530, CER of 0.099, WER of 0.180, BLEU score of 0.116, and ROUGE score of 0.505, indicating that the system correctly inferred and described the visual content. By incorporating Artificial Intelligence and assistive technology, diagrams for visually impaired students were easier to understand, interpret, and access within inclusive learning environments.

Overall, the study demonstrated that Artificial Intelligence (AI) extends the capabilities of traditional assistive technology beyond text reading into the perception of visual information. The proposed JAIF framework contributes to inclusive education, more autonomy for learners, and reduced barriers to access in STEM learning environments, especially when resources are limited.

5.2. Contributions of the Study

This research provides important contributions in academia, technology, and practice in assistive technology, Artificial Intelligence, and inclusive education. The study developed the JAWS–AI Integrated Assistive Framework (JAIF), which combines deep learning-based object detection, Optical Character Recognition (OCR), Natural Language Processing (NLP), and Text-to-Speech (TTS) technologies into an intelligent assistive pipeline interpreting educational diagrams and visual learning content.

Further, the research contributes technically through the development of a YOLOv12-based object detection model specifically constructed for structured STEM diagrams that include arrows, symbols, axes, electrical components, and graphic elements commonly used in science and technical education materials. This is intended to enhance the detection and understanding of visual content under controlled laboratory conditions. The research also implements an OCR and semantic interpretation framework that converts visual and textual information into context-aware descriptions rather than object labels and provides meaningful auditory explanations for visually impaired students.

On top of that, the research shows how Artificial Intelligence can be embedded within JAWS screen reader, expanding capabilities of assistive technology beyond the ability to read text to understand the visuals and the educational content being read. It represents an important step forward in AI-assisted accessibility systems for STEM education.

In addition, this study promotes inclusive education by providing independent learning opportunities for visually impaired students in STEM. The proposed system was alleviated barriers to accessibility of diagram-based instruction and promote equal learning in inclusive settings. Finally, the study provides a foundation for research and development in AI-powered assistive technology for schools and limited resources that have no other options to achieve advanced accessibility.

5.3. Limitations of the Study

While the proposed framework achieved promising results, there were several limitations of the proposed framework. First, the dataset used in training and evaluation was rather small and domain

specific, since the system was trained on STEM diagrams based on physics-related topics such as electrical circuits and free-body diagrams. This may limit its generalizability to other academic fields or visual data with complicated semantic information.

Second, the experiment was controlled. In the real world, poor image quality, varying lighting, blurred diagrams, annotation problem due to the software required free and noisy backgrounds may adversely affect object recognition and OCR. These conditions could affect the accuracy and reliability of the system during use.

Third, the platform is currently only equipped to process English-language educational content. Without multilingual and local language support, the system cannot be used for users who use their native Ethiopian languages as a means of learning and communicating. This lack of support for multiple languages does not allow the platform to be utilized in a variety of educational environments.

Fourth, though JAWS screen reader was integrated successfully, the frame relies on assistive software functionality. This dependence limits the ability to customize and directly improve the behavior of screen readers within the assistive software pipeline.

Fifth, the computational performance of the framework is dependent on the availability of sufficient hardware resources. Real-time processing may be reduced in low-resource education settings where computing and processing resources are limited. This may impact the scalability and practical application of the framework within resource constrained institutions.

Finally, this study was focused primarily on technical system testing using quantitative performance measures such as precision, recall, mAP, CER, WER, BLEU, and ROUGE. No large-scale user-centered usability test was conducted with visually impaired students in the classroom, so the evaluation of user experience, usability, and educational value is limited.

5.4. Recommendations for Future Work

Future research should focus on improving the practicality, accessibility, and generalization capability of the proposed JAWS–AI Integrated Assistive Technology Framework (JAIF). First, user-centered evaluation should be conducted with visually impaired learners and accessibility

experts to assess usability, task completion, comprehension, cognitive workload, perceived usefulness, and overall user satisfaction using standardized instruments such as the System Usability Scale (SUS).

Second, the educational effectiveness of JAIF should be investigated by comparing students' understanding of STEM diagrams when using JAIF with traditional screen readers such as JAWS alone, manual alternative text descriptions, and tactile graphics.

Third, robustness testing should be expanded using more challenging visual conditions, including low-resolution images, handwritten labels, rotated diagrams, noisy backgrounds, and previously unseen STEM diagram categories such as chemistry, biology, and advanced mathematics.

Fourth, future versions of the framework should support multilingual and localized accessibility by integrating Amharic and Afaan Oromo OCR and Text-to-Speech (TTS) capabilities, considering language encoding, voice availability, and translation quality.

Fifth, future research should explore integration with open-source screen readers such as NVDA to provide a lower-cost and more accessible alternative to JAWS, particularly for resource-constrained educational institutions.

Additionally, the framework should incorporate uncertainty communication mechanisms that provide confidence scores or warning messages when detections are uncertain, reducing the risk of presenting misleading information to visually impaired users.

Future improvements should also ensure that the JAIF interface itself is fully accessible by supporting keyboard navigation, screen-reader-friendly controls, accessible upload workflows, generated description history, and audio playback controls.

Finally, future studies should improve reproducibility by providing publicly accessible code repositories, model configurations, trained weights, sample datasets, annotation guidelines, and detailed instructions for deploying and evaluating the JAIF prototype.

References

[1] E. Volker, S. Gupta, and B. Brown, “Inclusive Education : Advantages and Overcoming Barriers,” vol. 6, no. 2021, pp. 1–6, 2022.

[2] U. Nations, “THE SALAMANCA STATEMENT FRAMEWORK,” no. June, pp. 7–10, 1994.

[3] T. Degener, “10 Years of Convention on the Rights of Persons With Disabilities,” *Netherlands Q. Hum. Rights*, vol. 35, no. 3, pp. 152–157, 2017, doi: 10.1177/0924051917722294.

[4] S. S. Review, “Inclusive Education : Strategies for Successful Inclusion of Students with Disabilities in Mainstream Classrooms,” vol. 4, no. 3, pp. 439–453, 2024.

[5] J. A. C. Castaneda *et al.*, “Designing inclusive tech playful educative solutions for visually impaired learners in STEM education,” *Smart Learn. Environ.*, 2025, doi: 10.1186/s40561-024-00358-x.

[6] F. Thea, F. H. Foundation, M. G. Foundation, L. Clubs, I. Foundation, and S. International, “Articles Trends in prevalence of blindness and distance and near vision impairment over 30 years : an analysis for the Global Burden of Disease Study,” no. 20, pp. 1–14, 2020, doi: 10.1016/S2214-109X(20)30425-3.

[7] K. Niraula, I. Journal, and E. S. Ijonses, “Exploring Experiences of Visually Impaired Youths in Tertiary Education To cite this article : Exploring Experiences of Visually Impaired Youths in Tertiary Education,” 2024.

[8] V. Impaired and S. Using, “Online and Biomedical Engineering,” vol. 21, no. 14, pp. 4–19, 2025.

[9] Z. I. Alam, S. Islam, and E. Hoque, “SeeChart: Enabling Accessible Visualizations Through Interactive Natural Language Interface For People with Visual Impairments,” *28th Int. Conf. Intell. User Interfaces (IUI ’23), March 27â•fi31, 2023, Sydney, NSW, Aust.*, vol. 1, no. 1, pp. 1–28, 2023, doi: 10.1145/3581641.3584099.

[10] A. Sharif, S. S. Chintalapati, J. O. Wobbrock, and K. Reinecke, “Understanding Screen-Reader Users ’ Experiences with Online Data Visualizations”.

[11] M. Mukhiddinov, “A Systematic Literature Review on the Automatic Creation of,” pp. 1–31, 2021.

[12] I. Haakma, M. J. Janssen, and A. E. M. G. Minnaert, “Need Support in Students with Visual Impairments: Comparing Teacher and Student Perspectives,” vol. 2, no. January, pp. 1–11, 2018, doi: 10.3389/feduc.2017.00071.

[13] T. E. Journal and H. E. Vol, “The Ethiopian Journal of Higher Education Vol. 7 No. 2 December 2020 1,” vol. 7, no. 2, pp. 1–45, 2025.

[14] T. H. E. Practices *et al.*, “COLLEGE OF EDUCATION AND BEHAVIORAL STUDIES,” 2024.

[15] T. G. Milki, B. H. Hailu, and A. Y. Male, “Capable but Excluded : Unpacking the Systemic Barriers to Mathematics Education for Students with Blindness,” vol. 05, no. 01, pp. 14–30, 2026.

[16] L. Orynbay, B. Razakhova, P. Peer, and B. Meden, “Recent Advances in Synthesis and Interaction of Speech , Text , and Vision,” pp. 1–25, 2024.

[17] R. Gonzalez, J. Collins, C. Tech, and C. Tech, “Investigating Use Cases of AI-Powered Scene Description Applications for Blind and Low Vision People,” 2024, doi: 10.1145/3613904.3642211.

[18] A. Khanal and R. Estrada, “Fully Automated Artery-Vein ratio and vascular tortuosity measurement in retinal fundus images,” 2023.

[19] D. Doermann, “YOLOv12: Attention-Centric Real-Time Object Detectors,” 2025.

[20] E. Dzierzgowska, M. Maćkowski, M. Kawulok, P. Brzoza, S. Maćkowska, and D. Spinczyk, “Alternative audio-graphic method for presenting structural information in mathematical graphs designed for low-vision users,” pp. 1–15, 2025.

[21] F. Escalona and M. Cazorla, “AIDEN : Design and Pilot Study of an AI Assistant for the Visually Impaired,” pp. 1–18, 2025.

[22] H. Miyauchi, “A systematic review on inclusive education of students with visual impairment,” *Educ. Sci.*, vol. 10, no. 11, pp. 1–15, 2020, doi: 10.3390/educsci10110346.

[23] A. Ghaffar, “ACADEMIA International Journal for Social Sciences ISSN-L (Online): 3006-6638 ACADEMIA International Journal for Social Sciences ISSN-L (Online): 3006-6638,” vol. 4, no. 3, pp. 1765–1774, 2025, doi: 10.63056/ACAD.004.03.0492.

[24] S. KIRBOYUN TİPİ, “How Screen Readers Impact the Academic Works of College and Graduate Students with Visual Impairments,” *Sak. Univ. J. Educ.*, vol. 13, no. 3, pp. 416–434, 2023, doi: 10.19126/suje.1201482.

[25] U. Oh, H. Joh, and Y. Lee, “Image Accessibility for Screen Reader Users : A Systematic Review and a Road Map,” 2021.

[26] M. Buji, M. Salminen, and J. Hamari, “International Journal of Human - Computer Studies Effects of Immersive Media on Emotion and Memory : An Experiment Comparing Article , 360-video , and Virtual Reality,” vol. 179, no. July, 2023, doi: 10.1016/j.ijhcs.2023.103118.

[27] W. Wang, B. Jing, X. Yu, Y. Sun, L. Yang, and C. Wang, “YOLO-OD : Obstacle Detection for Visually Impaired Navigation Assistance,” 2024.

[28] F. Elavsky, L. Nadolskis, and D. Moritz, “Data Navigator : An Accessibility-Centered Data Navigation Toolkit”.

[29] I. Technology *et al.*, “Screen Reader AI : A Conversational Web-Accessibility Assistant for Blind and Low-Vision Users,” vol. 5, no. 3, pp. 671–676, 2025.

[30] P. Naayini, P. K. Myakala, C. Bura, A. K. Jonnalagadda, and S. Kamatala, “AI-Powered Assistive Technologies for Visual Impairment,” 2025, [Online]. Available: http://arxiv.org/abs/2503.15494

[31] J. Sayyad and K. Attarde, “A systematic literature review on deep learning approaches for small object detection,” *Array*, vol. 29, no. November 2025, p. 100615, 2026, doi: 10.1016/j.array.2025.100615.

[32] S. A. Nawaz, J. Li, U. A. Bhatti, M. U. Shoukat, and R. M. Ahmad, “AI-based object detection latest trends in remote sensing , multimedia and agriculture applications,” no. November, pp. 1–21, 2022, doi: 10.3389/fpls.2022.1041514.

[33] S. Ren, K. He, R. Girshick, and J. Sun, “Faster R-CNN : Towards Real-Time Object Detection with Region Proposal Networks,” pp. 1–14.

[34] M. Hasan, K. M. Rabby, I. Jahan, J. A. Soeb, and F. Jubayer, “The Evolution and Advancement of YOLO Algorithms in Object Detection : From Real-Time Breakthroughs to Modern Architectures The Evolution and Advancement of YOLO Algorithms in Object Detection : From Real-Time Breakthroughs to Modern Architectures,” pp. 0–31, 2025, doi: 10.20944/preprints202510.2019.v1.

[35] S. A. Adelakun, “Exploring STEM Kit Diagrams for Braille Readers in Inclusive Classrooms,” pp. 1–10.

[36] Z. J. Muhsin, R. Qahwaji, and F. Ghanchi, “Review of substitutive assistive tools and technologies for people with visual impairments : recent advancements and prospects,” pp. 135–156, 2024.

[37] V. Priyadharsini and R. S. Mary, “Universal Design for Learning (UDL) in Inclusive Education : Accelerating Learning for All,” pp. 145–150, 2024.

[38] C. Kooli and R. Chakraoui, “AI-driven assistive technologies in inclusive education : benefits , challenges , and policy recommendations,” *Sustain. Futur.*, vol. 10, no. July, p. 101042, 2025, doi: 10.1016/j.sftr.2025.101042.

[39] F. Momotaz and S. M. Billah, “Understanding the Usages , Lifecycle , and Opportunities of Screen Readers ’ Plugins,” vol. 16, no. 2, 2023, doi: 10.1145/3582697.

[40] I. Students, “JAWS (Job Access With Speech): A Third Eye for Visually”, doi: 10.4018/978-1-7998-4736-6.ch002.

[41] A. Surve, H. Nikam, S. Pandey, J. Santani, and V. Pujari, “Vision-Aid : AI Powered Assistive System for the Visually Impaired with Advanced Speech and Object Recognition,” vol. 10, no. 9, pp. 150–155, 2025.

[42] K. Srinivasan, S. V. V, G. Vignesh, and R. Subash, “Object Recognition for Visually Impaired People,” vol. 68, no. 8, pp. 33–38, 2020.

[43] K. Zhang, Y. Wang, S. Shi, Q. Wang, C. Wang, and S. Liu, “Improved yolov5 algorithm combined with depth camera and embedded system for blind indoor visual assistance,” pp. 1–15, 2024.

[44] C. V Jul, C. Science, and H. Hd, “YOLO V 5 , YOLO V 8 AND YOLO V 10 : T HE G O -T O D ETECTORS FOR R EAL - TIME V ISION,” pp. 1–12, 2024.

[45] A. Wang *et al.*, “YOLOv10 : Real-Time End-to-End Object Detection,” no. NeurIPS, pp. 1–28, 2024.

[46] R. Deshpande, “Handwritten formula recognition using CNN,” vol. 5, no. 1, pp. 95–98, 2024.

[47] H. Park, M. Sung, and D. Kim, “Real-Time On-Device Continual Learning Based on a Combined Nearest Class Mean and Replay Method for Smartphone Gesture Recognition,” pp. 1–19, 2025.

[48] C. Yan, H. Hutter, F. M. Schmitt-koopmann, and A. Darvishy, “Chart Accessibility : A Review of Current Alt Text Generation,” *IEEE Access*, vol. PP, p. 1, 2025, doi: 10.1109/ACCESS.2025.3571626.

[49] Q. Gao, R. Manduchi, and G. Legge, “VI-OCR : " Visually Impaired " Optical Character Recognition Pipeline for Text Accessibility Assessment VI-OCR : " Visually Impaired " Optical Character Recognition Pipeline for Text Accessibility Assessment,” pp. 0–18, 2025.

[50] A. T. Aleme, “RIGHT TO INCLUSIVE EDUCATION OF PERSONS WITH DISABILITIES :,” *Tadesse*, 2023.

[51] B. Li *et al.*, “arXiv : 2207 . 11463v1 [cs . CV] 23 Jul 2022 When Counting Meets HMER :,” 2022.

[52] R. Arifando and S. Eto, “applied sciences Improved YOLOv5-Based Lightweight Object Detection Algorithm for People with Visual Impairment to Detect Buses,” 2023.

[53] B. Kim, M. A. Alawami, E. Kim, S. Oh, J. Park, and H. Kim, “A Comparative Study of Time Series Anomaly Detection Models for Industrial Control Systems,” pp. 1–21, 2023.

[54] L. Korada, V. K. Sikha, and D. Siramgari, “INTELLIGENT SYSTEMS AND APPLICATIONS IN AI & Accessibility : A Conceptual Framework for Inclusive Technology,” vol. 12, pp. 983–992, 2024.

[55] A. Wang *et al.*, “YOLOv10 : Real-Time End-to-End Object Detection,” no. NeurIPS, pp. 1–21, 2024.

[56] “A Comprehensive Study on Image Preprocessing Techniques for Enhanced Image Analysis Noise Reduction”.

[57] M. B. Laili, M. Z. Abdulah, E. Sunardi, and J. Permatasari, “Integrating YOLOv8 , EasyOCR , and GTTS for text detection in assistive technology for the visually impaired,” vol. 2, pp. 1–8, 2025, doi: 10.31603/bistycs.185.

[58] S. Bhushan, M. Lee, S. Korea, and S. Korea, “Block Diagram-to-Text : Understanding Block Diagram Images by Generating Natural Language Descriptors,” pp. 153–168, 2022.

[59] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, and S. Gould, “Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering,” *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 6077–6086, 2018, doi: 10.1109/CVPR.2018.00636.

[60] D. Rizky, N. Herawati, and E. P. Sartinah, “The Role of Text To Speech Assistive Technology to Improve Reading Ability in E-Learning for ADHD Students,” vol. 6, no. 2, pp. 169–174, 2018.

[61] G. Waters, “AI testing , evaluation , verification and validation for accessibility : a comprehensive framework,” no. February, pp. 1–19, 2026, doi: 10.3389/fdgth.2025.1679603.

[62] H. Fahmi and G. Sahuri, “Artificial intelligence multilingual image-to-speech for accessibility and text recognition,” vol. 14, no. 3, pp. 1743–1751, 2025, doi: 10.11591/ijai.v14.i3.pp1743-1751.