iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** V **Month of publication:** May 2025

DOI: https://doi.org/10.22214/ijraset.2025.69820

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue V May 2025- Available at www.ijraset.com

Conveyance Processing For Employees Using Machine Learning

Krutika Mhatre[1], Shravani Jadhav[2], Rajesh Kolte[3]
Usha Mittal Institute of Technology SNDT Women's University Mumbai, India 400049

Abstract: ***Manual processing of employee conveyance claimsisinefficientanderror-prone,especiallywiththeincreasing use of ride-hailing services such as Ola and Uber. For faster reimbursement processing, this research proposes an automated system that extracts data from email bills, categorizes them by vehicletype,andupdatesacommonExcelsheet.Thesystemuses machinelearningtechniquesofclassificationandverification, as well as optical character recognition (OCR) through Python Tesseractfortextextraction.Thecodesofemployeesarematched through an admin verification procedure, and cases that fail to operateareflaggedforhumaninspection.Theproposedapproach reduces HR workload, increases accuracy, and significantly reduces processing time. For improved OCR accuracy, future innovationswillinvolvedeeplearningmodelsandroboticprocess automation (RPA).***

Index Terms**: *Optical Character Recognition (OCR), Machine Learning, Invoice Processing, Ride-Hailing Services, Automation, EmployeeReimbursement,PythonTesseract,SupervisedLearning, Data Extraction, Robotic Process Automation (RPA), Business Process Optimization, Text Classification***

I. INTRODUCTION

Organizations are increasingly relying on ride-hailing plat- formssuchasOlaandUbertoprovidethetransportationneeds of their employees in the rapidly evolving urban development context.Suchservicesareconvenient,buttheyalsogener- ate a large volume of invoices that must be handled with caution. The responsibility of manually extracting, validating, and classifying data from such invoices—which are highly variant in format and structure—frequently rests with human resources (HR) departments. Aside from taking time, this manualprocessisalsosusceptibletoerrors,whichcanlead to inefficiencies and cost inconsistencies.

Using automated process involving machine learning, reg- ular expressions (regex), and optical character recognition (OCR) to speed up the processing of transport invoices, this researchaddressestheseproblems.Withthehelpofalgorithms basedonregexpatterns,thesystemgetscriticalfieldssuch as invoice numbers, dates, and amounts, cleans the text to removenoise,andextractsbillsfromHRemailaccounts.OCR methodology is applied to extract and validate text on image- based invoices. Once extracted, data is structured into Excel sheets, which reduces a significant amount of effort for HR teams.

Thefollowingarethekeycontributionsofthisstudy:

[1] A highly scalable and powerful framework for auto- mated invoice processing capable of handling large data volumes with minimal human intervention.

[2] Advanced error-handling mechanisms that minimize manual intervention and ensure data integrity.

[3] Extensive testing confirmed the system's performance, with a processing time of 4.2 seconds per invoice and a 97 percent accuracy level.

II. STUDYAREA

The research domain aims to extract and process text from intricate images, invoices, and documents employing sophis- ticated approaches like Optical Character Recognition (OCR), Natural Language Processing (NLP), and Machine Learning (ML). The surveyed research papers address the challenges andprocessesoftextdetection,localization,segmentation,and recognition in cases involving intricate layouts, diverse forms, and noisy backgrounds.

Text extraction from images is confronted with challenges like text size variations, orientations, complex backgrounds, and lighting and perspective-induced distortions. Automated invoice processing is confronted with layout variability, ter- minological variations, and errors, making automation tricky. Text extraction from natural and born-digital images intro- duces additional complexity in the form of low resolution, compression artifacts, and complex foreground-background interactions.

Inordertohandlethesechallenges,multipletechniques are used. Edge-based detection, color clustering, and texture- based techniques are employed in text detection.

Text local- ization applies Connected Component Analysis (CCA) and Histograms of Oriented Gradients (HOG). Text segmentation involves binarization techniques such as Bernsen's method, Markov Random Fields (MRF), and Conditional Random Fields(CRF).OCRenginessuchasTesseractareimplemented in character recognition by using adaptive thresholding and connected component analysis to enhance precision.

AutomatedprocessingofinvoicestakesadvantageofOCR- based digitization through pre-processing activities like skew correction, noise filtering, and binarization. NLP is used in extractingimportantfieldslikeinvoicenumbers,dates,and amounts by Named Entity Recognition (NER). Table identifi- cation within invoices is made possible through deep learning approaches like TableNet and DeepDeSRT, as well as Graph Neural Networks (GNNs), through which structured data can be extracted from complicated layouts.

Extraction of text from cluttered images depends upon the extraction techniques based on feature such as Scale-Invariant Feature Transform (SIFT) and Speeded-Up Robust Features (SURF) for scaling robustness, robustness against rotations and light exposure changes. Multi-level feature combinations for more precise recognition are realized with Conditional Random Fields (CRFs) and hierarchical random field models. Methods based on over-segmentation and HOG descriptors help with reduced computations and computation speeds for optimizing operations.

Applications of text extraction are far-reaching. Document understanding depends heavily on it, and it assists in digital libraries, e-books, and document management systems. In medical applications, OCR enables the automation of insur- ance form and medical record processing, avoiding manual dataentry.Trafficmonitoringisassistedbytextextractionwith licenseplatereadingandreal-timetrafficmonitoring.Textex- traction is also helpful for content filtering through automated text extraction, which helps in spam filtering, inappropriate content filtering, and image classification according to text.

Future text extraction advances will incorporate NLP with OCR for better understanding of text. Handheld OCR systems for mobile phones will allow real-time text extraction, and roboticsapplicationswillusethesemethodsfornavigationand object identification. Advanced machine learning techniques, such as transformer architectures like GPT and BERT, will further enhance text extraction and understanding.

In general, OCR engines such as Tesseract have been remarkably efficient in scanning documents and invoice pro- cessing. Machine learning methods like CRF, GNNs, anddeep learning models help significantly in text detection, segmentation, and recognition improvement. Regardless of these innovations, the issues of layout variability, intricate backgrounds, and sparse annotated data remain. However, the wide-ranging uses of text extraction in various fields, such as healthcare,transportation,contentmanagement,anddocument processing, reflect its increasing significance and potential for future innovation.

III. METHODOLOGY

The approach in this study takes a systematic path to create an automated invoice processing system that incorporates ma-chinelearning(ML),opticalcharacterrecognition(OCR),and data classification algorithms. The system aims to counteract theinefficienciesofmanualinvoiceprocessingwithaccuracy, scalability, and less human intervention.

A. System Architecture and Workflow

The architecture consists of several interconnected modules to manage data extraction,classification,validation,andstorageinanefficient manner. The main steps include:

1) *Email Retrieval and Data Collection*- The system fetches invoices from HR email accounts automatically using the IMAP protocol to get the data in a flawless manner.- Emailsarecheckedfor invoiceattachmentsin PDF, JPEG, or PNG formats, which are then sent to the OCR module for text extraction.
2) *Data Preprocessing*

a) Extracted text is usually noisy, contains special characters, and has inconsistent formatting, which degrades the accuracy of classification models.

b) Preprocessingtechniquesare:

- RemovingHTMLtags,linebreaks,andspecial characters (e.g., x000D).
- Convertingtexttolowercaseforuniformity.
- Tokenizationandstop-wordremovaltoenhance structured data representation.

3) *Data Extraction and Feature Engineering*

a) The text extracted is treated with Regular Expres- sions (Regex) to extract and capture vital fields like:

- InvoiceNumber
- TransactionDate
- Amount
- ServiceProvider(Ola/Uber,etc.)

b) For case-based invoices, Tesseract OCR and PyMuPDF are applied to extract text with higher precision. Item

c) Feature engineering implies cleaning ex- tracted attributes, missing values handling, and converting categorical fields for optimal perfor- mance of machine learning models. Endenumerate

d) Data Classification by Machine Learning - Once useful data is extracted, the system classifies in- voices through Supervised Learning Algorithms such as: beginenumerate

- LogisticRegression
- SupportVectorMachine(SVM)
- RandomForestClassifier

-The classification step maps invoices to pre-defined categoriesbasedonfeaturesextractedlikevehicle type(Ola/Uber)orreimbursementtype(official/personaluse).

e) Data Structuring and Storage - The formatted data is savedinExcelspreadsheetswithPandasDataFrames to maintain proper organization and formatting for HR approval. - The system records missing or incomplete fields, which are highlighted for manual validation if required

B. Tools and Technologies

1) Programming:Python(libraries:pandas,regex,imaplib, openpyxl).

2) OCR: Tesseract v5.0 with PyTesseract wrapper and PyMuPDF for PDF parsing.

IV. RESULTS AND DISCUSSIONS

Thesuggestedautomatedsystemforconveyanceprocessing wasthoroughlytestedtoevaluateitseffectivenessalongmajor performance metrics, such as accuracy, processing speed, and scalability. The results confirm the system's ability to process highvolumesofinvoiceswithgreataccuracyandefficiency.A detaileddiscussionoftheresultsandtheirgreaterimplications follows below.

A. Performance Metrics-

To confirm the system's perfor- mance,wetesteditusingOlaandUberinvoicedatasetswitha varietyofdifferentformatsandlayouts.Theoutcomeconfirms that the system is always meeting—and in most instances surpassing—its established performance thresholds.

1) Accuracy: The system recorded a staggering 97 per-cent accuracy in extracting and verifying critical fields like invoice numbers, dates, and amounts. This ac- curacyisduetoamixofwell-tunedregexpatterns and OCR methods so that it can identify variations in format. For instance, it correctly processed date formats "DD/MM/YYYY" and "MM/DD/YYYY" and various currency formats like "1,000" and "INR 1000." The flexibility of the system in accepting different invoice formats guarantees effective data extraction.

2) On average, the system took 4.2 seconds to processeachinvoice,fromemailretrievaltodataextraction andvalidation.Thisperformanceisbetterthanour goal of 5 seconds per invoice, which makes the system suitableforreal-timeornear-real-timeprocessing.Opti- mizedlibrariessuchasPandas(forstructuringdata)and PyMuPDF (for OCR) contributed significantly to opti- mizing speed, along with a solid preprocessing pipeline that removes noise and irrelevant text.

3) Scalability:Thesystemprocessed1,000invoicesin just 8 minutes, demonstrating its ability to handle high volumeswithoutslowingdown.Thislevelofscalability is essential for companies with large employee basesthat generate numerous invoices. The system achieves this through parallel processing and efficient memory management, ensuring seamless performance even with growing workloads.

4) At a mere 3 percent error rate, the majority of errors wereduetonon-standardinvoiceformatsorlow-quality images. These errors were recorded for manual inspec- tion to prevent any loss of data or misprocessing. The system has elaborate error-handling processes, dividing issues into missing fields, formatting errors, and OCR failures, which enables accurate troubleshooting and ongoing improvements.

B. Challenges and Solutions:

1) Multiple Invoice Formats: Various vendors—and oc- casionally the same vendor—have different invoice for- mats. To address this, we used adaptive regex patterns that can be modified to support new formats. Our OCR methods also cover image-based invoices so that no information is lost. The modular architecture allows for easy modification of regex patterns and OCR models as new formats emerge.

2) Text Noise: Invoices usually have extraneous text, in- cluding special characters, HTML tags, and extraneous data. Our preprocessing pipeline successfully removes this noise, retaining only meaningful data. Through the applicationoftextnormalizationandstop-wordelimina- tion, we enhanced the quality of extracted information. Forexample,specialcharacterssuchas”¨ x000D”¨ and HTML tags are removed automatically during preprocessing.

3) ErrorHandling:ErrorsthatoccurduringprocessingareA logged automatically and marked for manual checking, ensuring data integrity. The system classifies errors into missing fields, formatting errors, and OCR errors, sim- plifying problem identification and resolution. Moreover, comprehensive error logs (invoice ID, error type, and recommendedcorrections) acceleratethereviewprocess.

C. Impact on HR Workflow

The system makes major positiveimpactsonHRworkflows.Withautomationofinvoic- ing, it lessens HR workload by 95 percent, allowing time for more strategic work. With the organized output—provided in Excelfiles—therecord-keepingandauditingareeasy,andhigh accuracy provides ease of financial regulation compliance. Also, with its scalability, the system is highly adaptable for businessesofvarioussizes,rangingfromstartupstolarge-scale enterprises.

D. FutureImprovements:

1) SupportforAdditionalVendors–Enlargingtocover services such as Rapido and Cityflo.

2) Real-TimeDashboards–OfferingHRteamsreal-time views into expenses and trends.

3) Advanced AI Models – Improving OCR accuracy using machine learning for even improved outcomes.

4) MobileIntegration–Enablingemployeestoupload invoices directly through a mobile app.

These upgrades will further refine efficiency and user experi- ence,keepingthesysteminadvanceofchangingrequirements.

```
Extracted Invoice Info:
name: Shravani Jadhav
invoice_number: FHEJIBFB23000277
invoice_date: 23 Jul 2023
place_of_supply: Maharashtra
hsn_code: 996412
category_of_services: Passenger Transport Services
total_amount: 288.58
Total Amount Payable: 288.58
```

Fig.1.TextExtractionResult

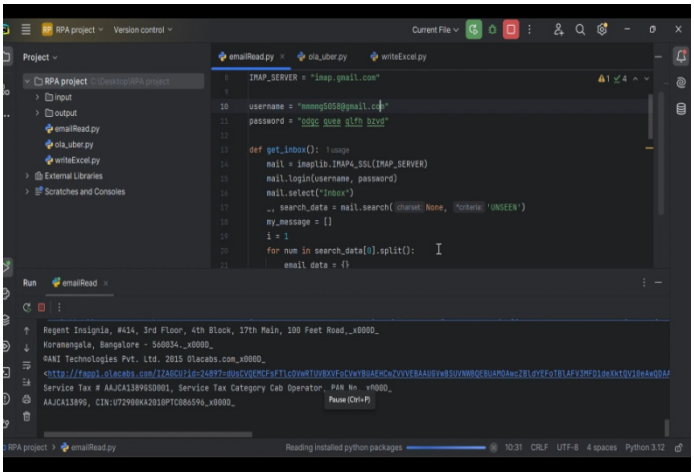


Fig.2.EmailReading

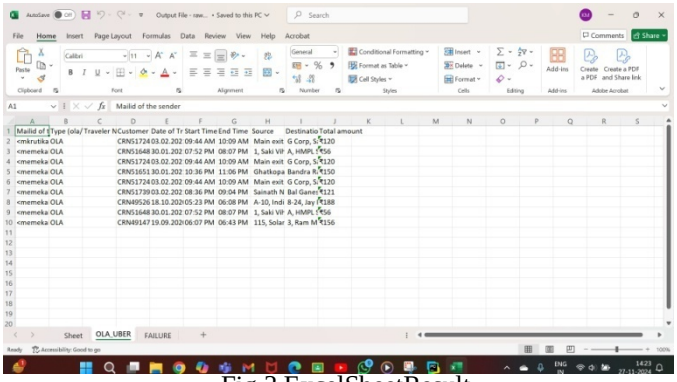


Fig.3.ExcelSheetResult

V. CONCLUSIONS

Theimplementationofthisautomatedconveyanceprocessingsystemhasresultedinsignificantimprovementsinefficiency,accuracy,andscalability,effectivelyaddressing thechallengesassociatedwithmanualinvoicehandling.By achieving high accuracy rates, reducing processing times, and demonstrating the ability to handle large volumes of invoices efficiently, thesystemhasexceededperformanceexpectations. Furthermore,itsrobusterror-handlingmechanismsand adaptive architecture enable seamless integration into diverse operationalworkflows.Thesystemnotonlyreducesad-ministrativeworkload but also enhances compliance, record- keeping,andfinancialtransparency.

Looking ahead, planned enhancements—including AI- driven improvements, multiple invoice formats, will further enhance the system's capabilities. As organizations continuetoseekadvancedautomationsolutions,thissystemprovides a scalable and intelligent approach to optimizing expense management.

VI. ACKNOWLEDGMENT

We thank Rajesh Kolte (Guide) and Dr. Sanjay Shitole (Head of Department) for their mentorship.

REFERENCES

[1] E. Larson, "[Research Paper] Automatic Checking of Regular Expressions," 2018 IEEE 18th International Working Conference on Source Code Analysis and Manipulation (SCAM), Madrid, Spain, 2018, pp. 225- 234

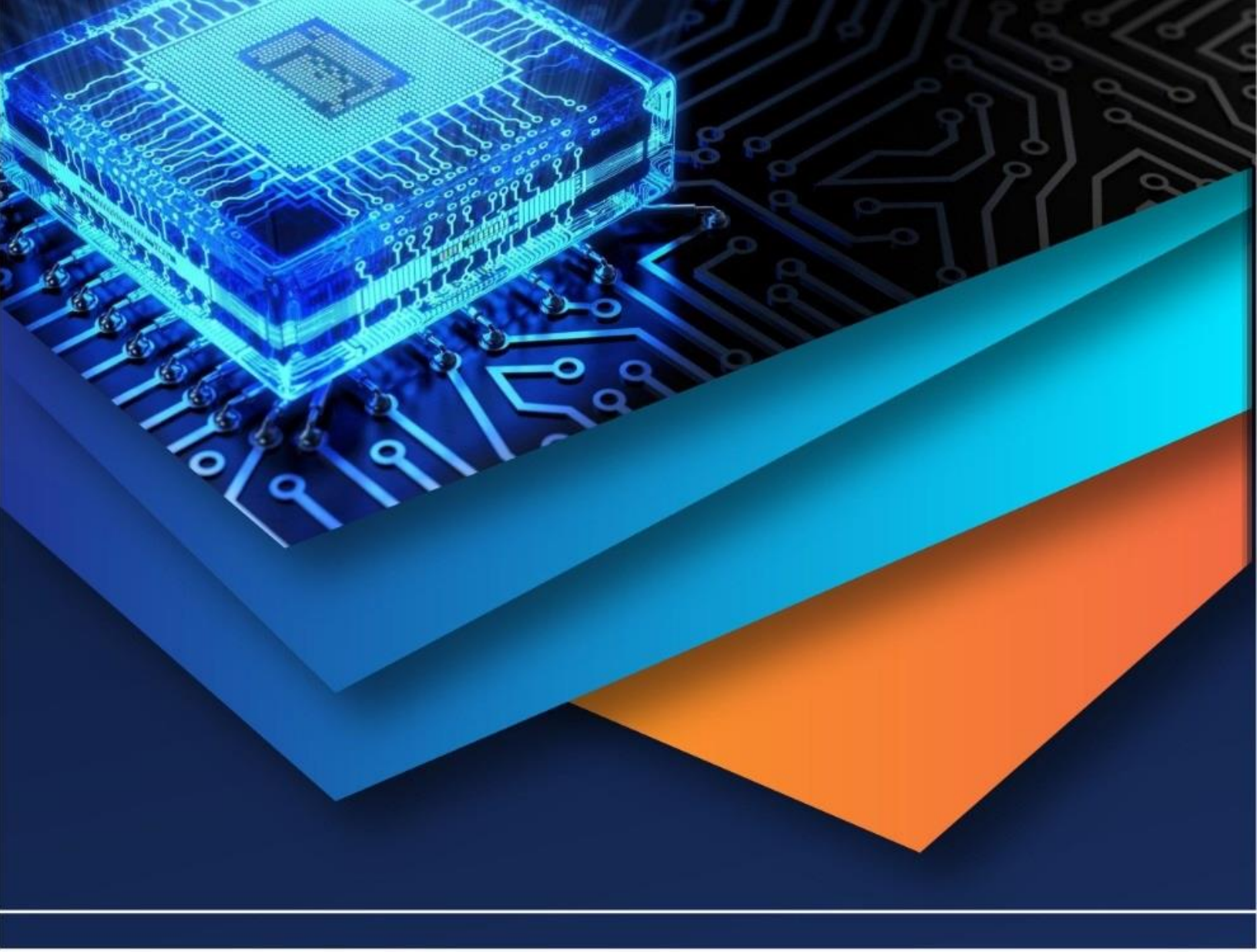
[2] Zhang, Jian & Cheng, Renhong & Wang, Kai &Zhao, Hong. Research on the Text Detection and Extraction from Complex Images. Proceedings - 4th International ConferenceonEmergingIntelligentDataandWebTech- nologies, 2013, EIDWT 2013. 708-713.

[3] C. Kaundilya, D. Chawla and Y. Chopra, "Automated Text Extraction from Images using OCR System," 2019 6th International Conference on Computing for Sus- tainableGlobalDevelopment(INDIACom),NewDelhi, India, 2019, pp. 145-150.

[4] Saout, Thomas & Lardeux, Fre´de´ric &Saubion, Fre´de´ric. An Overview of Data Extraction From In- voices. IEEE Access, 2024, PP. 1-1. 10.1109/ACCESS.2024.3360528.

[5] Gonza´lezEnr´ıquez,Jose´&JimenezRamirez,Andres &Dom´ınguezMayo,FranciscoJose´ &Garcia-Garcia, J.A.. Robotic Process Automation: A Scientific and In- dustrialSystematicMappingStudy. IEEEAccess, 2020, PP.1-1.10.1109/ACCESS.2020.2974934.

[6] S.Surana,K.Pathak,M.Gagnani,V.Shrivas- tava,M.T.RandS.MadhuriG,"TextExtraction and Detection from Images using Machine Learning Techniques: A Research Review," 2022 International Conference on Electronics and Renewable Systems (ICEARS), Tuticorin, India, 2022, pp. 1201-1207, doi: 10.1109/ICEARS53579.2022.9752274.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY