

Volume 2. No. 2
December 2024

E-ISSN- 3048-5460
Subscribers copy
Not for sale



i-manager's

Journal on Data Science & Big Data Analytics

Extracting Intelligence from Data



i-manager's Journal on Data Science & Big Data Analytics

About the Journal

i-manager's Journal on Data Science & Big Data Analytics aims to bring the best of research and innovations in the field of Data Science and Engineering and reveal the hidden patterns found in the data captured in very large databases with the precision of human intelligence. The Journal will bring bright ideas ignited in the young minds that will explore future opportunities in the area of data security, data intelligence, business intelligence, supply chain automation, industrial automation and fields that are yet to be envisioned. The Journal aims to provide a platform for academicians, researchers and students to expose the knowledge from the evolving domain of data science with fresh reviews, research experiment with new concepts in data modelling, case studies of successes in implementation of new technologies etc.

i-manager's Journal on Data Science & Big Data Analytics is published by i-manager Publications, one of India's leading Academic Journal Publisher, publishing 48 Academic Journals in diverse fields of Engineering, Education, Management and Science.

Why Publish with us

i-manager Publications currently publishes academic Journals in Education, Engineering, Scientific and Management streams. All of i-manager's Journals are supported by highly qualified Editorial Board members who help in presenting high quality content issue after issue. We follow stringent Double Blind Peer Review process to maintain the high quality of our Journals. Our Journals target both Indian as well as International researchers and serve as a medium for knowledge transfer between the developed and developing countries. The Journals have a good mix of International and Indian academic contributions, with the peer-review committee set up with International Educators.

Submission Procedure

First time authors are requested to register themselves on the registration page in the website. Once registered, authors can submit and know the status of their papers.

Researchers and practitioners are expected to submit on or before the stipulated deadline. Authors of accepted proposals will be notified before the stipulated deadline about the status of their proposals. All submitted articles in full text are expected to be submitted before the stipulated deadline.

Review Procedure

All submissions will undergo an abstract review and a double blind review on the full papers.

The abstracts would be reviewed initially and the acceptance and rejection of the abstracts would be notified to the corresponding authors. Once the authors submit the full papers in accordance to the suggestions in the abstract review report, the papers would be forwarded for double-blind review by the review panel. The final selection of the papers would be based on the report of the review panel members, and the approval of the Editor-in-Chief.

Format for Citing Papers

Author surname, initials (s.) (2024). Title of paper. i-manager's Journal on Data Science & Big Data Analytics, 2(2), xx-xx.

Copyright

Copyright © i-manager Publications 2024. All rights reserved. No part of this Journal may be reproduced in any form without permission in writing from the publisher.

Contact e-mails

*editor@imanagerjournals.com
submissions@imanagerpublications.com*

i-manager's Journal on Data Science & Big Data Analytics

Editor-in-Chief

Dr. R. A. Thakur
Director IITM,
India.

EDITORIAL COMMITTEE

Dr. S. Padmaja	Head of the Department, Department of Computer Science & Engineering, Keshav Memorial Institute of Technology, Hyderabad, Telangana, India.	Dr. Prakash G. L.	Associate Professor, Department of Information Science and Engineering, BMS Institute of Technology, Visvesvaraya Technological University, Bangalore, Karnataka, India.
Dr. Sonia Rathee	Assistant Professor, Department of Computer Science and Engineering, Maharaja Surajmal Institute of Technology, New Delhi, India.	Dr. T. S. Ravi Kiran	Associate Professor & Head, Department of Computer Science, P. B. Siddhartha College of Arts & Science, Vijayawada, Andhra Pradesh, India.
Dr. Suraj Shah	Assistant Professor, Ganpat University - Centre for Management Studies and Research, Gujarat, India.	Dr. M. Anline Rejula	Visiting Lecturer, Department of Computer Applications-PG, Scott Christian College (Autonomous), Nagercoil, Tamil Nadu, India.
Ms. Kandula Neha	Assistant Professor, Department of Computer Science and Engineering, BVRIT Hyderabad College of Engineering for Women, Hyderabad, Telangana, India.	Dr. Ayush Gautam	Analytics Advisory Specialist, Accenture, Gurugram, Haryana, India.

Abstracting / Indexing



i-manager's Journal on Data Science & Big Data Analytics

OUR TEAM

Publisher

Joe Winston

Renisha Winston
Editorial Director

Sahaya Nijuba S.
Associate Editor

Varsha M.
Assistant Editor

Merlin J.
Assistant Editor

Suganya G.
Assistant Editor

Anitha Bennet
GM - Sales & Marketing

Bibin D William
Design Head

M. Sajintha
Senior Designer

Shejin P. M.
Assistant Designer

S. Ravina
Journal Metadata Executive

OUR OFFICES

Registered Office

3/343, Hill view,
Town Railway Nager,
Nagercoil, Kanyakumari District - 629001
Ph : 91-4652- 277675
E-mail : info@imanagerpublications.com

Editorial Office

13-B, Popular Building,
Mead Street, College Road,
Nagercoil, Kanyakumari District - 629001
Ph : (91-4652) 231675, 232675, 276675
E-mail : editor@imanagerjournals.com

Join with us



<https://www.facebook.com/profile.php?id=61552769060011>



<https://www.facebook.com/imanagerPublishing/>



<https://twitter.com/imanagerpub>

CONTENTS

RESEARCH PAPERS

- | | |
|----|---|
| 1 | BREAST CANCER DIAGNOSIS MODEL BASED ON CONVOLUTIONAL NEURAL NETWORKS' MULTIPLE ARCHITECTURES
By Ayobami Ekundayo, Enesi Femi Aminu, Winifred Omone Saliu, Oluwaseun Adeniyi Ojerinde, Uchenna Cosmas Ugwuoke |
| 20 | IMPACT OF ARTIFICIAL INTELLIGENCE ON CYBER SHOPPING IN KANNIYAKUMARI DISTRICT
By Goldie Sheila Jesolif C. S., Chithra James |
| 28 | AN INTERACTIVE VISUALIZATION AND DATA ANALYSIS OF SUPERMARKET STORE
By Likita H. L., Hiresh Yadav, Anjali Hedao, Swagota Bera, Siddhartha Choubey |
| 35 | ROLE OF ARTIFICIAL INTELLIGENCE IN INVESTMENT MANAGEMENT
By Mahiba G. B., Jesintha P. |
| 40 | A GENERATIVE AI MODEL FOR FOREST FIRE PREDICTION AND DETECTION
By Nallusamy M., Karthick S., Vetrivel B. |

EDITORIAL

The current issue of *i-manager's Journal on Data Science & Big Data Analytics (JDS)*, (December 2024: Volume 2 Issue 2) has five peer-reviewed papers that presents some interesting research on Data Science and Big Data Analytics.

Ekundayo et al. explores a breast cancer diagnosis model leveraging multiple architectures of Convolutional Neural Networks (CNNs). The rise of artificial intelligence and machine learning, particularly CNNs, offers significant potential for addressing the challenge of early breast cancer detection. This study aims to identify the CNN architecture that delivers optimal performance in terms of accuracy, F1 score, and Cohen Kappa score. The evaluation focuses on Custom Optimized CNN, ResNet, and EfficientNet architectures, with ResNet demonstrating superior performance across all five evaluated metrics.

Goldie and Chithra explore the influence of AI on cybershoppers in Kanniyakumari District, Tamil Nadu, India. This study collected primary data using a convenience sampling method with a sample size of 200 participants. Age emerged as a significant factor, highlighting distinct preferences and comfort levels with AI technologies between younger and older consumers. Among AI-based online shopping platforms, Amazon was the most popular, while Indiamart saw the least usage. This study underscores the importance of demographic factors in shaping customer preferences and engagement with online shopping platforms.

Likita et al. introduce an advanced interactive visualization system designed for analyzing data within a supermarket environment. An intelligent model was developed using linear regression, LASSO regression, and XGBoost, outperforming traditional models. This study incorporates data pre-processing, feature engineering, and algorithmic improvements to provide insights into customer behavior, sales trends, and inventory management efficiency. This study seeks to enhance decision-making in inventory management, marketing strategies, and sales forecasting by leveraging data-driven insights to improve the efficiency and effectiveness of supermarket operations.

Mahiba and Jesintha investigate the most preferred AI investing apps for investment management and assess investor awareness of AI-driven investment strategies. The study involved 120 investors from Kanniyakumari District, selected using the convenience sampling method. Both primary and secondary data were collected, and SPSS tools were utilized for data analysis. This study highlights AI's transformative impact on investment management, improving decision-making, optimizing portfolios, and delivering personalized strategies. This ultimately fosters more efficient and effective investment outcomes in today's complex financial landscape.

Nallusamy et al. present a novel lightweight fire detection approach utilizing Deep Convolutional Neural Networks (DCNN) with a focus on temporal aspects to enhance accuracy. The study aims to advance forest fire detection, addressing the destructive impact of wildfires on ecosystems and communities. Unlike traditional methods that rely on handcrafted features such as color, motion, and texture, which struggle with accuracy, this study offers a significant improvement. By leveraging DCNN, it provides a promising solution to the ongoing challenge of timely and reliable forest fire detection.

We thank the authors for their research work, and reviewers for their time in reviewing the papers. Our special thanks are due to the Editor-in-Chief Dr. R. A. Thakur for his support in enhancing the quality of the Journal.

Hope this issue imparts an enlightening reading experience! Enjoy Reading!

Warm regards,

Sahaya Nijuba S.
Associate Editor
i-manager Publications

BREAST CANCER DIAGNOSIS MODEL BASED ON CONVOLUTIONAL NEURAL NETWORKS' MULTIPLE ARCHITECTURES

By

AYOBAMI EKUNDAYO *

ENESI FEMI AMINU **

WINIFRED OMONE SALIU ***

OLUWASEUN ADENIYI OJERINDE ****

UCHENNA COSMAS UGWUOKE *****

*-***** Department of Computer Science, Federal University of Technology, Minna, Nigeria.

<https://doi.org/10.26634/jds.2.2.20921>

Date Received: 13/06/2024

Date Revised: 02/07/2024

Date Accepted: 11/07/2024

ABSTRACT

In 2020, the World Health Organization (WHO) estimated that 2.3 million women worldwide were diagnosed with breast cancer, which resulted in 685,000 deaths. According to the projections, the number of women who have been diagnosed with breast cancer over the last five years before and by the end of 2020 was expected to reach 7.8 million, making it the most common type of cancer worldwide. Early diagnosis could prevent the ailment, however, lack of availability of health facilities and the cost of accessing treatment, especially in developing nations, are among the challenges confronting the solution. With the advent of artificial intelligence and machine learning models, specifically Convolutional Neural Networks (CNNs), considering their multiple architectures is highly promising to address the challenge of early diagnosis. Therefore, this study aims to propose an architecture of CNNs that gives the best accuracy, F1 score, and Cohen Kappa score among the Custom Optimized CNN, ResNet, and EfficientNet architectures. From the results, ResNet's performance across the five metrics outweighs the other two architectures. While ResNet reported an accuracy, precision, and F1 score of 0.9987, 0.9934, and 0.9950, respectively, EfficientNet, which has the second performance, reported 0.9977, 0.9914, and 0.9939 as accuracy, precision, and F1 score, respectively. Therefore, the best-performing architecture can be deployed for other available breast cancer datasets in order to ensure its total efficiency.

Keywords: Breast Cancer Diagnosis Model, Early Diagnosis, Custom Optimized CNN, ResNet, EfficientNet, Multiple Architectures, Artificial Intelligence, Machine Learning Models.

INTRODUCTION

The World Health Organization (WHO) estimates that 2.3 million women globally were diagnosed with breast cancer in 2020, leading to 685,000 deaths. By the end of 2020, the total number of women diagnosed with breast cancer will have reached 7.8 million, making it the most

prevalent cancer worldwide (Samee et al., 2022). The primary cause of cancer-related mortality for women is breast cancer, which accounted for 11.6%. GLOBOCAN reported that over six hundred and twenty-five thousand people died from breast cancer in 2018, out of 2.09 million new cases that were identified. This disease, which causes cells to grow abnormally, affects women worldwide. It can start in the lobules, which are the milk-producing glands; the ducts, which are the tubes that carry milk to the connective tissue, which is the fatty and fibrous tissue that surrounds and holds everything together



This paper has objectives related to SDGs



(Abdelli et al., 2020). Approximately 23% of all malignancies in women worldwide are breast cancers, making them the most frequent cancer among women. In the West, between one in eight and nine women will develop breast cancer at some point in their lifetime (Yurttakal et al., 2020). Breast cancer is ranked second worldwide in terms of cancer-related deaths among women and is the most diagnosed malignancy (Xie et al., 2020). Breast cancer is classified according to the growth, size, and stage of the tumor cell. Bowel and malignant tumors are the two main types of breast cancer, the former are nonspreading tumors, while the latter are more invasive, life-threatening, and complex than benign tumors (Shivhare & Saxena, 2021). Early diagnosis with immediate precautions and treatments can avert the potential danger associated with the ailment. However, due to the high cost of accessing medical facilities and the unavailability of these facilities, it is difficult for early diagnosis. With the advancement in technology, some top-notch techniques of artificial intelligence (AI) are promising to provide alternative, cheaper, and more accessible facilities for early diagnosis. In this development, soft computing and machine learning (ML) techniques are the foundation of computer-aided intelligent and expert diagnostic systems, which are indispensable (critical) instruments for assessing and forecasting breast cancer and can support oncologists (medical professionals) in their decision-making (Vijayakumar et al., 2021).

Furthermore, Deep Learning (DL) models are a class of machine learning techniques that delegate feature extraction and classification to the machine. The DL models have demonstrated outstanding performance in picture classification, even though the feature extracted can be applied to other applications (Oyelade & Ezugwu, 2021). Examples of these models included DL+ Transfer Learning (TL), DL GoogleNet+VGG, DL+TL, and the proposed DCNN (Raza et al., 2023). In medical imaging, Convolutional Neural Networks (CNNs) have improved tumor detection and classification efficiency over classical methods (Houssein et al., 2022). CNN maintains the spatial integrity of mammograms, including the way

pixels are connected to generate unique features, despite training difficulties brought on by a dearth of labeled images (Mahmood et al., 2022). Therefore, the goal of this study is to determine the best architecture for precise breast cancer diagnosis by evaluating the performance of these several CNN models, which include EfficientNet, ResNet, and the Custom Optimized CNN. The findings of this study have the potential to further the field of medical image analysis and support the current breast cancer research endeavor.

1. Related Works

Das et al. (2023) evaluated and compared the effectiveness of pre-trained deep convolutional neural network designs trained on mammography images with the suggested shallow convolutional neural network architecture with varying specifications. The initial attempt in the study to do the automatic identification of BC involves preprocessing of mammogram images. The final data is then fed into three distinct kinds of shallow convolutional neural networks with representational differences. In the second method, pre-trained convolutional neural networks VGG19, ResNet50, MobileNet-v2, Inceptionv3, Xception, and Inception-ResNet-v2 are fed the same collection of images through transfer learning via fine tuning. The accuracy for the CBIS-DDSM and INbreast datasets in the experiment with two datasets is 80.4%, 89.2%, and 87.8%, 95.1%, respectively. Khan et al. (2021) constructed a Deep Convolutional Neural Network (CNN) in order to segment and categorize the kinds of breast abnormalities, including calcifications, masses, asymmetry, and carcinomas. The pre-trained model ResNet50 was used to do transfer learning on the dataset. The study created an improved deep learning model where learning rate is regarded as one of the key factors in neural network training. In the suggested approach, the learning rate is adaptively set according to changes in error curves during the learning process. The suggested deep learning model has classified these four categories of abnormalities, such as masses, calcifications, carcinomas, and asymmetry, such as mammograms, with an 88% accuracy.

Computer-assisted methods based on deep learning can analyze even minor pathologies in detail, which could advance treatment for a comprehensive evaluation. Altan (2020) proposed a CNN-based DL model that sought to improve the DL models' ability to learn for an operational breast cancer diagnosis while also evaluating the suitability of several feature-learning models. In order to assess the classification performances in accordance with different CNN designs, the mammograms were fed into the DL. High classification performance rates of 92.84%, 95.30%, and 96.72% were attained by the suggested Deep model in terms of accuracy, sensitivity, specificity, and precision, respectively.

Xie et al. (2020) suggested an automated multi-scale end-to-end deep neural network model for mammogram classification, which only requires mammogram images and class labels (without ROI annotations). The suggested model produced three feature map scales that enable the classifier to combine global information with local lesions for classification. Additionally, the images processed have a lower percentage of non-breast pixels and preserve as much information about small lesions as possible, which helps the model to concentrate on the small lesions. Based on the features of mammography images, the BRS module is suggested to preprocess the raw mammogram images. The INbreast dataset is used to verify the performance of our method. The proposed model works better when compared to other cutting-edge mammography classification algorithms. Additionally, networks with fewer parameters that can achieve equivalent performance and save 60% of the computational resources are deployed with the multi-scale technique. Mammography is a useful tool for diagnosing breast cancer. However, the accuracy of the results is primarily dependent on radiologists' clinical backgrounds. Zhang et al. (2020) used a convolutional neural network to do a two-stage classification (normal/ abnormal and benign/ malignancy) of two-view mammograms. In order to classify mammograms from two viewpoints, the study built a multi-view feature fusion network model.

Additionally, it suggested using a multi-scale attention DenseNet as the main network for feature extraction. To extract the features of two mammograms taken from different angles, the model has two independent branches. The construction of the multi-scale convolution module and attention module is the main goal of this study. The model performed well in both classification tasks, according to the final experimental data. The suggested approach was assessed using the DDSM database. The classification of normal and abnormal mammograms has accuracy, sensitivity, and AUC values of 94.92%, 96.52%, and 94.72%, respectively. Additionally, the classification of benign and malignant mammography had accuracy, sensitivity, and AUC values of 95.24%, 96.11%, and 95.03%, respectively.

Yektaei et al. (2019) aimed at diagnosing breast cancer using multiscale CNN. A Multiscale Convolutional Neural Network (MCNN) method is presented for the classification of tumors. Based on the MCNN architecture, which feeds many deep CNNs with varying sizes and resolutions of mammography pictures, the traditional phase of manually crafting feature extraction is omitted. The website has labeled images, a total of 1024 samples from 322 people, including those with benign and aggressive breast cancer, were labeled in this study. In this dataset, 322 samples of cancer patients' profiles were gathered in 1994. The suggested method allows for a safer computer-aided diagnosis of pleural cancer since it has higher classification rates than the traditional state-of-the-art techniques. Using the multiscale convolution methodology, the study achieved a diagnosis accuracy of 97%, revealing the effective suggested method.

Tasnim et al. (2021) classified breast cancer images using multiple convolutional neural network architectures. To handle breast cell imaging data, convolutional neural network (CNN) models are developed. Xception, GoogLeNet, VGG19, InceptionV3, and AlexNet models are employed in the categorization of cells that are either non-invasive ductal carcinoma (Non-IDC) or invasive ductal carcinoma (IDC). The learning rate is recorded by testing and training the models at various epochs. The study showed that accuracy rises and data loss

decreases with increasing epochs. InceptionV3 and Xception have accuracy rates of 92.48% and 90.72%, respectively. Similar to this, the accuracy of VGG19 and AlexNet is 94.83% and 96.74%, respectively. With an accuracy of 97.80%, GoogLeNet outperforms the other developed models, nonetheless. The GoogLeNet model detects the IDC cells that cause breast cancer with a high degree of accuracy and precision.

Samala et al. (2018) aimed at developing a deep Convolutional Neural Network (CNN) with a multistage transfer learning strategy that uses information from related auxiliary domains for intermediate-stage fine-tuning in order to distinguish between benign and malignant masses in Digital Breast Tomosynthesis (DBT), Data on breast imaging from DBT, Digital Mammography (DM), and digitized Screen-Film Mammography (SFM) were obtained, resulting in 4,039 distinct ROIs, 2,242 benign, and 1,797 malignant. By adjusting the degree to which the convolutional layers were frozen, the optimal transfer network was chosen through cross-validation among six transfer networks. CNN knowledge that had been trained on ImageNet data was immediately refined using DBT data in a one-stage transfer learning method. Using a multi-stage transfer learning approach, the DBT data was used to fine-tune acquired information from ImageNet with mammography data. For the second step of transfer learning, the effectiveness of freezing the entire CNN structure versus the first convolutional layer was evaluated between two transfer networks. It was investigated how the size of the training sample affected the classification performance for different transfer learning and fine-tuning strategies by changing the training data from 1% to 100% of the sets that were available. A performance metric that was employed was the area under the receiver operating characteristic curve (AUC). For single-stage transfer learning, the view-based AUC on the test set was 0.85_0.05; for multi-stage learning, it improved significantly ($p < 0.05$) to 0.91_0.03.

The diagnosis of Architectural Distortion (AD) in mammograms is difficult because of its tiny size and subtle, variable asymmetry on the breast mass. Nonetheless, Rehman et al. (2021) predicted breast

cancer from digital mammograms, an automated computer-aided diagnosis approach based on architectural distortion utilizing computer vision and deep learning was developed. Four phases are included in the proposed mammography classification framework: picture preprocessing, image augmentation, and image pixel-wise segmentation. To categorize AD's ROIs into benign and malignant groups, machine learning networks, deep learning, and architectural distortion ROI detection are used. Using computer vision and depth-wise 2D V-net 64 convolutional neural networks, the suggested method has been assessed on three databases: the PINUM, the CBIS-DDSM, and the DDSM mammography images. Accuracy values of 0.95, 0.97, and 0.98 were obtained, respectively. According to experimental results, the suggested approach performed better than ShuffleNet, MobileNet, SVM, K-NN, and RF. To reduce the overhead of manual analysis, Eren and Tarhan (2020) introduced Convolutional Neural Networks (CNNs) for breast mass identification. The Region of Interest (ROI) and Region Proposal Network (RPN) sections of the faster R-CNN are modified for the automatic identification of breast mass abnormalities in the CNN architecture, which is intended for the feature extraction stage. The algorithm distinguished between benign and malignant abnormalities in Mammography (MG) pictures based on the detection of mass regions. MG pictures from various local hospitals were gathered for the suggested model. The images underwent various preprocessing processes, including bilateral, median, and gaussian filters, to remove the breast region from the MG image's background. The model's performance on the test dataset is determined to be 91.86% for detection accuracy, 94.67% for sensitivity, and 92.2% for AUC-ROC.

Breast cancer patients' chances of survival can be significantly increased by early identification. Niu et al. (2021) primarily used the convolutional neural network method to distinguish between benign and malignant breast tumors in mammography. To extract the features of global and local image patches, the study used densely connected networks and multi-scale residual networks as the backbone networks. Second, the study optimized the

two feature extraction networks using the Convolutional Block Attention Module (CBAM) attention module to increase the network's feature expression capability. In order to classify benign and malignant breast masses, the study finally merged the features of multi-scale picture patches. In the DDSM database, the accuracy, sensitivity, AUC value, and corresponding standard deviation of the method are 0.9626 ± 0.0110 , 0.9719 ± 0.0126 , and 0.9576 ± 0.0064 , respectively. Compared with the commonly used ResNet (AUC = 0.8823 ± 0.0112) and DenseNet (AUC = 0.9141 ± 0.0085), the performance of the method has improved.

Many efforts have been made in the form of screening programs for prevention, as breast cancer remains one of the top causes of death for women. In addition to classifying breast cancers as benign or malignant, Nawaz et al. (2018) predicted which subclasses of tumors, such as lobular carcinoma and fibroadenoma, would arise. When compared to state-of-the-art models, experimental results on histopathology images using the BreakHis dataset demonstrate that the DenseNet CNN model obtained high processing performances with 95.4% percent accuracy in the multi-class breast cancer classification test. One crucial step towards diagnosis and treatment follow-up in medical imaging is the segmentation of breast tumors. Radiologists can lessen the heavy manual workload associated with breast cancer analysis by automating this difficult operation. However, by creating two fully Convolutional Neural Networks (CNN) based on SegNet and U-Net, El Adoui et al. (2019) suggested two deep learning techniques to automate breast tumor segmentation in Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI). In this study, models were trained and validated on a dataset of 86 DCE-MRIs of 43 patients with local advanced breast cancer, obtained both before and after two cycles of chemotherapy. A total of 5452 slices were used in the study. A high-performance architecture made of graphic processing units was employed in order to shorten the training period. On 85% and 15% of the data, respectively, the model was trained and verified. SegNet produced a mean intersection over union (IoU) of

68.88, while U-Net produced an IoU of 76.14%.

In order to improve the automatic detection of breast cancer, Alanazi et al. (2021) proposed a Convolutional Neural Network (CNN) method in the study. The method analyzed hostile ductal carcinoma tissue zones in Whole-Slide Images (WSIs). The suggested system uses different CNN architectures to automatically detect breast cancer. Three (LR, KNN, and SVM) machine learning models were used in the study. The dataset used in the study was obtained from Kaggle 162 H&E. The data set consists of both benign and malignant images. A large collection of around 275,000 RGB image patches measuring 50 by 50 pixels served as the basis for all architectures. Quantitative results underwent validation testing utilizing each methodology's performance metrics. The suggested method is proven to be effective, yielding outcomes with 87% accuracy, which may lessen human error throughout the diagnosing procedure. Apart from helping to advance several domains like drug development, time-series modeling, and optimization techniques, deep learning algorithms have produced outstanding outcomes in medical diagnosis and picture analysis. Burçak et al. (2021) suggested a deep convolution neural network model. The model computes the initial weight of the network and updates the model parameters for quicker backpropagation learning using a variety of methods, including stochastic gradient descent, Nesterov accelerated gradient, adaptive gradient, RMSprop, AdaDelta, and Adam. The study employed a parallel computing architecture with a graphics processing unit that was Cuda-enabled in order to train the model quickly and with less hardware. The outcomes showed that the deep convolutional neural network model performs well in classification, reaching an accuracy score of 99.05%.

2. Proposed Methodology

The proposed methodology evaluates and assesses the performance of the architectures based on the available dataset, which is derived from the MIAS Mammography Dataset obtained from the Kaggle repository and contains 322 high-resolution mammographic images from approximately 161 patients. Figure 1 shows an

overview of the proposed diagnostic model within a conceptual framework. A comprehensive design and implementation guide supports decisions related to hardware, software, and network infrastructure. The guiding principles, patterns, and connections in the high-level design ensure scalability, maintainability, reliability, and performance of the components.

A well-coordinated flow through numerous interconnected elements creates an efficient breast cancer diagnosis model in the proposed model's framework as shown in Figure 1. The design identifies various components and the interaction among these components, such as the user, the Data Treatment panel, the Feature Engineering panel and the Model Training

panel. The functions of each design element are clearly stated as:

- The "user," interacting with the system, initiates the acquisition of data from the MIAS dataset available on "Kaggle." This is the first step toward executing further activities in this architecture.
- The data is processed through the "Data Treatment Panel," a sophisticated function that ensures all obtained data complies with quality requirements and is still relevant. A number of steps are involved in the process, such as "Data Cleaning," which fixes errors or inaccuracies, "Data Normalization," which streamlines inputs into uniform units for processing, and, finally, "Data Augmentation," which increases

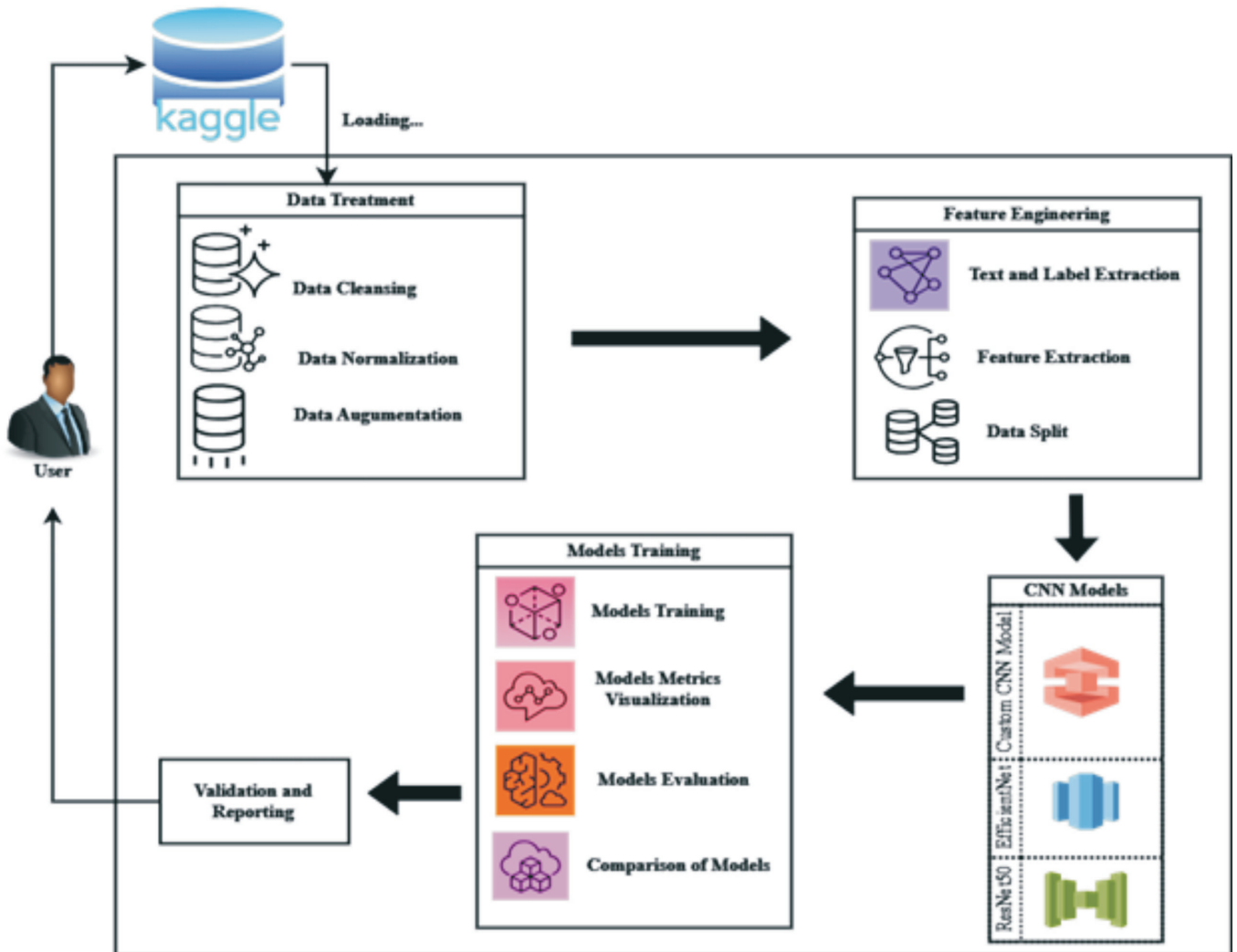


Figure 1. Conceptual Framework of the Breast Cancer Diagnosis Model

diversity by increasing dataset size with the goal of strengthening the model.

- Following the first round of data processing, the information is then sent to the "Feature Engineering Panel." Relevant information is extracted using a variety of methods, including "Text and Label Extraction," which parses textual material. Important features from the input dataset are also found through a process called "Feature Extraction." In addition, there is a stage called "Data Split," where datasets are divided into training, validation, and testing sets.
- The journey reaches the "CNN Models Panel," which has three distinct Convolutional Neural Network (CNN) models: an arbitrary CNN, ResNet, and EfficientNet. Each model's architecture is defined prior to training commencing in the "Model Training Panel."
- In addition to providing metrics that monitor the models' development and performance, the training dataset teaches the models in the "Model Training Panel." This phase is divided into three parts: "Models Training," which gets the model ready for tasks ahead, "Models Metrics Visualization," which assesses the model's effectiveness, and "Models Evaluation," which looks at the trained model using a different set to evaluate it. The final step compares the results of ResNet, EfficientNet, and custom-optimized CNN.
- The "Validation and Reporting" phase is critical since it entails validating the selected model to guarantee its dependability and efficacy in practical settings. This stage makes sure that performance is carefully assessed before measurements, findings, and insights are shared with users. The ultimate objective is to provide a comprehensive understanding of the competency capacities of the models for breast cancer diagnosis that are in use.

Figure 2 shows the flow of information from input sources through several processing steps to produce the final.

The data flow diagram of the proposed model, as shown in Figure 2, starts with data source identification, using the MIAS mammography dataset as the primary resource.

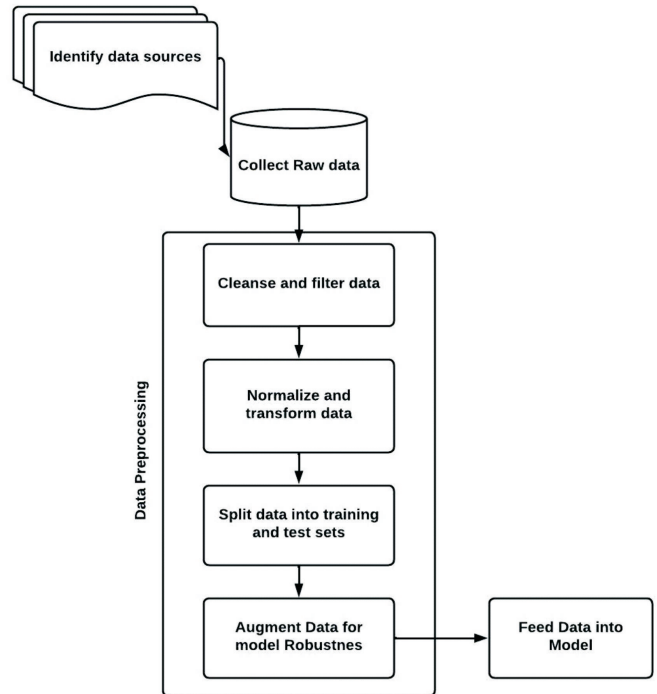


Figure 2. Data Flow Diagram for the Proposed Model

Then, raw data (a collection of mammography images) is obtained, and the data preprocessing stage entails multiple complex steps targeted at improving the dataset. These complex phases include the removal of noise by filtering and cleaning, normalization of pixel values, scaling to ensure consistency, and data separation to separate training and test sets. Image rotation is one of the approaches used to improve dataset variety, which strengthens the resilience of the model. The last phase involves feeding carefully preprocessed and enhanced data into several CNN designs, including ResNet, a specially designed model, and EfficientNet. The refined features taken from the preprocessed images in the training dataset are used to train each architecture. By following a methodical procedure, the model gets exposed to a large and consistent dataset, which improves its ability to diagnose breast cancer. The emphasis on data preprocessing not only smooths the input data but also places the model in a way that maximizes performance in a variety of circumstances. Table 1 shows the algorithmic framework for the proposed multiple CNN architectures.

The algorithm for creating the breast cancer diagnosis

Algorithm for the Breast Cancer Diagnosis Model using Multiple CNN Architectures

Input: MIAS Mammography Dataset

Output: Breast Cancer Diagnosis Model

Parameters: MIAS mammography dataset (t), the image data (i), ResNet model (r), Custom Optimized Model (c), EfficientNet model (e), data augmentation (da), training (tg), validation (vn), testing sets (ts), training progress (tp), accuracy (a), precision (p), recall (r), F1-score (Fs), breast cancer diagnosis model (bcdm), evaluation metrics (em), best-performing model (bfm).

Procedure:

1. Input t
2. Preprocess t by resizing l
3. Apply da by rotating l.
4. Split i into tg, vn, and ts.
5. Assign labels to i, 0 for benign, 1 for malignant.
6. Choose : r, c and e
7. Define the architecture of r, c and e
8. Train r, c and e using the training dataset.
9. Monitor tp and vn performance for all the models.
10. Evaluate the trained CNN models on the test dataset.
11. Calculate classification metrics of a, p, r, Fs
12. Compare the performance of all the CNN models.
13. Select the bfm based on em.
14. Return bcdm.

Table 1. Algorithmic Framework of the Proposed Model

model is shown in Table 1. It begins by utilizing the MIAS mammography dataset (t) as input. All of the photos in this set are then preprocessed and resized to 224 x 224 pixels, which is a common size. This study uses data augmentation techniques to add rotation to images in the training data to increase their variety. These altered photos will be divided into three sets: a training set, a validation set, and a testing set after the pre-processing and augmentation procedures. Based on the results of their diagnosis, each and every one of them has been given a label, with an "0" denoting benign instances and an "1" denoting malignant ones. The algorithm chooses three Convolutional Neural Network (CNN) architectures: ResNet (r), a custom-built model (c), and EfficientNet (e). Each architecture's particular layers and links are specified. When this is finished, the training dataset is used to monitor progress through validation performance for each of the three CNN architectures that were chosen: Resnet (r), Custom Model (c), and EfficientNet (e) in order to prevent overfitting during this process.

The trained CNN models are tested on new, unseen data using a different test dataset. A number of classification metrics, including accuracy, precision, recall, and F1-score, are computed to quantitatively assess each

model's performance. Three CNN models are compared and tested for performance, and the algorithm finally chooses the best-performing model based on evaluation metrics, concluding with the finalizing of the Breast Cancer Diagnosis Model development process.

The creation of CNN's architecture is the initial stage of model creation. Considering the intricacy of mammography pictures and the requirement for maximum precision, a deep CNN architecture is utilized. A pre-trained model, like ResNet, is used as the foundation architecture through the use of transfer learning. The pre-trained model's convolutional layers will be used to extract features, and the fully connected layers will be tailored to the specific binary classification job of determining whether breast cancer is benign or malignant.

3. Results and Discussion

This study uses a number of frameworks and software packages extensively. Python is the most widely used programming language for having a large library and strong community in data science and machine learning. Deep learning models are created with TensorFlow and Keras, two Python-based tools. Image data can be read, altered, and enhanced with the help of OpenCV, another essential library for image processing. To ensure optimal performance, one of the models' hyperparameters is optimized using Keras Tuner. Furthermore, tools and libraries like Seaborn, Numpy, and Pandas make it easier to manipulate, analyze, and visualize data.

3.1 Data Visualization and Preprocessing

The machine learning begins with an understanding and preparation of the MIAS mammography dataset, which this study obtained from the Kaggle repository. The diagnostic algorithm was trained using mammography pictures from this dataset, which is a major source of information for breast cancer research. An initial look at the features of the mammography images is provided by visualizing the data, which also highlights any obvious patterns or anomalies. Figure 3 shows the differences in tumor presentations and overall image quality by

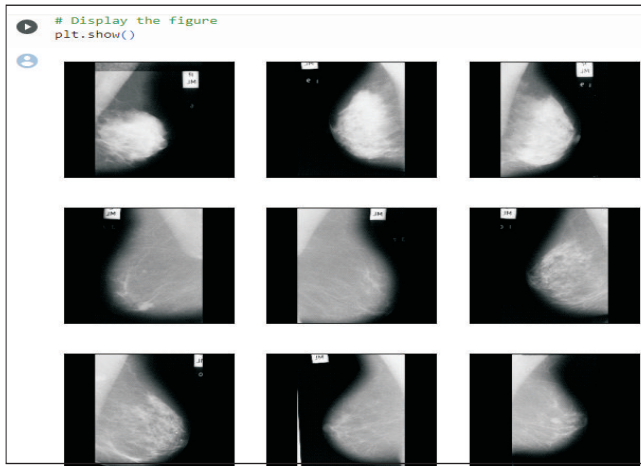


Figure 3. Snapshot of Data Visualization Montage from the Notebook

presenting a subset of the MIAS dataset photos.

3.2 Image Rotation for Data Augmentation

CNNs in particular are deep learning models that do well with enormous volumes of data. The MIAS dataset is small, hence, methods of data augmentation, like image rotation, are used to make the dataset appear larger than it actually is. Changing the angle at which images are rotated adds useful diversity, which enables the model to learn from different representations and may improve its generalization abilities.

Figure 4 shows the images following the augmentation procedure. Rotation matrices for a 2-degree angle are created using the `getmatrix` function. A loop iterates 180 times, using these matrices to rotate each image in turn. The dataset is greatly increased by this technique, which yields 180 rotated variants of each original image. The goal of the augmented dataset is to give the model a

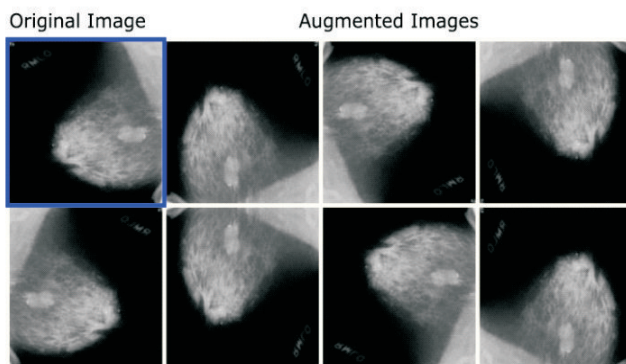


Figure 4. Visual Representation of the Data Augmentation Process

more diverse set of training examples by methodically changing the orientation of the images. This will eventually increase the number of images in the dataset from 322 to 57,960 images.

3.3 Labels Processing

Each image in the MIAS dataset is assigned a label indicating its diagnostic classification as either benign or malignant. Extracting and processing these labels is essential to ensure the model receives the correct supervisory signals during training. This phase involves reading the label annotations and matching them to the corresponding images.

Figure 5 shows the function's code snippet for label extraction. The 'Info.txt' file contains information that can be read and processed by the `read_label` function. Labels in this file identify whether a mammography image is malignant (cancerous) or benign (non-cancerous).

3.4 Data Splitting

The data is divided into training and testing subsets after being supplemented, displayed, and labeled. This segregation ensures that our model is trained on one part of the data (the training set) and evaluated on a completely separate subset (the testing set), allowing for an accurate assessment of the model's diagnostic performance, as shown in Figure 6. Figure 6 shows the dataset distribution: 80% for training and 20% for testing.

3.5 Model Building and Hyperparameter Optimization

In any machine learning system, the model's architecture and hyperparameters are crucial in determining how well the model performs, particularly when utilizing deep learning approaches like Convolutional Neural Networks (CNNs). The architecture outlines the layers, their kinds, and the connections between them and the overall structure of the model. Conversely, the training process is controlled by hyperparameters, such as the learning rate for the optimizer or the number of filters in a convolutional layer. They are pre-set, not learned from the data, and their optimization can have a big impact on the model's performance.

3.5.1 Convolutional Neural Network (CNN) Architecture

CNNs are very good at classifying images because they

```
[ ] # Function to Read labels from file
def read_label():
    """Read labels from file."""
    print("Reading labels")
    filename = url + 'Info.txt'
    text_all = open(filename).read()
    lines = text_all.split('\n')
    info = {} # Dictionary for label data

    for line in lines:
        words = line.split(' ')
        if len(words) > 3:
            if (words[3] == 'B'): # Label 'B' for benign
                info[words[0]] = {angle: 0 for angle in range(0, no_angles, 2)} # Assigning label 0 for benign
            if (words[3] == 'M'): # Label 'M' for malignant
                info[words[0]] = {angle: 1 for angle in range(0, no_angles, 2)} # Assigning label 1 for malignant

    return info
```

Figure 5. Code Snippet of the Function for the Label Extraction

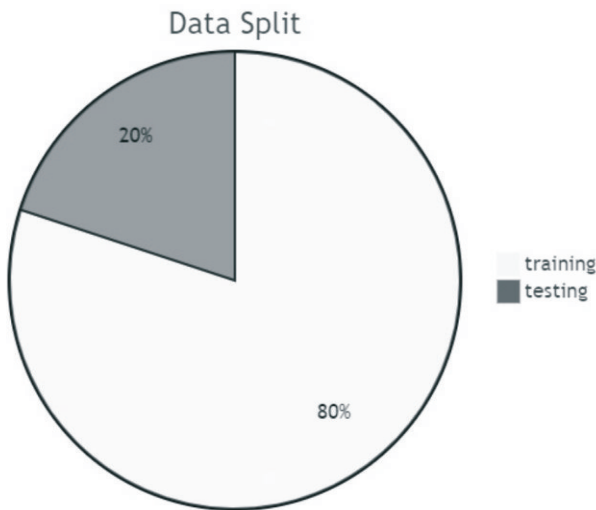


Figure 6. Distribution of Dataset

can capture the spatial hierarchy of features. The model utilized for this study has multiple convolutional layers, each of which is followed by pooling layers and activation functions. With deeper layers identifying more intricate structures and earlier layers identifying simpler patterns like edges, these layers seek to extract characteristics and patterns from the images. The network's last layers, which are thick layers, aid in the classification process by identifying if a picture shows benign or malignant tumors. The construction of a Convolutional Neural Network (CNN) was achieved using Keras. Layers are stacked in a sequential manner because the sequential class is used

to initialize the model.

- *Input Layer:* The input shape corresponds to the dimensions of the images (rows × columns × channels) with a single channel, indicating grayscale images.
- *Convolutional Layers:* The first layer has 32 filters of size 3 × 3 × 3, followed by a ReLU activation function. The subsequent layers increase the number of filters, moving from 64 to 128 and then decreasing back to 64 and 32. Each of these layers also uses 3 × 3 × 3 filters and is activated using the ReLU function.
- *Pooling Layers:* After each convolutional layer, there's a max-pooling layer with a pool size of 2 × 2 × 2. This layer downsamples the feature maps, reducing their dimensions by half.
- *Dropout Layers:* After the pooling operation in the later convolutional layers, dropout layers are added. These layers randomly set a fraction (20% in this case) of the input units to 0 at each update during training, which helps prevent overfitting.
- *Flattening Layer:* After the convolutional and pooling operations, the 2D feature maps are flattened into a 1D vector. This is a prerequisite before feeding the data into a fully connected layer.
- *Dense Layer:* A dense layer with a single neuron is

added, which outputs the prediction. This layer uses the sigmoid activation function, making it suitable for binary classification tasks.

3.5.1.1 Hyperparameter Optimization Using Keras Tuner

Hyperparameters regulate the finer points of the training procedure, while the model architecture provides the overall framework. Nevertheless, choosing the ideal hyperparameters by hand might be a time-consuming and inefficient method. As such, instruments such as Keras Tuner are utilized. The procedure is automated by Keras Tuner, which looks through a variety of hyperparameter settings to choose the set that provides the greatest performance. The RandomSearch function of the Keras Tuner was used for this study as it selects hyperparameters at random from predetermined ranges and uses them to train the model. This method produces findings that are reliable and efficient in terms of computing. The model building of the CNN is created using a method named `build_model`. A sequential model is initialized by carrying out actions that are:

- Two convolutional layers are added. For these layers, the number of filters and the size of the filters (referred to as the kernel size) are determined by hyperparameters that will be optimized.
- After each convolutional layer, an activation function (ReLU) and a pooling layer are added to reduce the spatial dimensions.
- A dropout layer is introduced after the second convolutional layer to prevent overfitting. The dropout rate is also a hyperparameter to be optimized.
- The model is flattened, and a dense layer with a sigmoid activation function is added for binary classification.
- The model is compiled with an optimizer, loss function, and evaluation metric. The learning rate for the optimizer is another hyperparameter to be optimized.

The RandomSearch function of the Keras Tuner is initialized. It builds models with various combinations of hyperparameters using the `build_model` function. The hyperparameter search spaces are:

- Number of filters for the convolutional layers.
- Kernel size for the first convolutional layer.
- Dropout rate after the second convolutional layer.
- Learning rate for the optimizer.

The RandomSearch method maximizes validation accuracy. It tries a maximum of 5 different hyperparameter combinations. For each combination, the model is trained three times to average out the performance. Using the training data, the search approach trains the models with various combinations of hyperparameters.

3.5.1.2 Optimal Hyperparameters

The optimal hyperparameters produced the highest validation accuracy after the hyperparameter optimization process. These ideal values reveal the preferences of the model, including the best learning rate or the quantity of filters to use in a convolutional layer. When these hyperparameters are used in full-scale training, the model is guaranteed to be trained as efficiently as possible, optimizing its potential accuracy on unobserved data. Table 2 shows the optimal hyperparameters derived from the search, where the Conv_1 Filter, Conv_2 Filter, Dropout_2, and Learning Rate values were 32, 48, 0.1, and 0.0001, respectively.

3.5.1.3 Model Training and Evaluation

The prepared dataset was used to train the convolutional neural network, which was optimized with the best hyperparameters. The diagnostic accuracy and robustness of the model were then assessed based on a variety of measures.

The model was trained on the dataset as an essential next step after defining the CNN architecture and figuring out the best hyperparameters. When a neural network is trained, input data is fed into it, and the model's internal weights are modified in response to the prediction error.

Hyperparameter	Optimal Value
Conv_1 Filter	32
Conv_2 Filter	48
Dropout_2	0.1
Learning Rate	0.0001

Table 2. Optimal Hyperparameters Derived from the Search

The model's goal over the 100 epochs of training was to reduce error, hence improving its capacity to identify breast cancer from mammograms. With 128 samples, the model changed its weights because the batch size was set to 128. This provides a balance between the frequency of model updates and computing efficiency. The optimized breast cancer diagnosis model is therefore trained and evaluated. The optimized model training was carried out based on the tasks, including:

- The model named `best_model`, which has the most suitable settings (hyperparameters) from previous steps, is trained.
- The training data (`x_train` and `y_train`) is used for this purpose.
- A portion (20%) of the training data is set aside for validation to monitor the model's performance during training.
- The training is done over 100 cycles (known as epochs).
- The dataset is shuffled in each epoch to ensure randomness.
- The model is trained in small batches of 128 images at a time, which helps in updating the model's internal settings more frequently.
- The `callbacks=[es]` part is a mechanism to monitor the training process and possibly stop early if the model isn't improving.

Once training is complete, the model's performance is tested on a separate set of data (`x_test` and `y_test`) that it hasn't seen before. The model's prediction error (`loss_value`) and accuracy (`accuracy`) on this test data are calculated and printed. The accuracy tells us how often the model's predictions were correct.

3.5.1.4 Model Evaluation Metrics

To determine the model's efficacy, it is essential to assess its performance on untrained data after training. To evaluate the diagnostic power of the model, several measures were applied. A precise indicator of the proportion of accurate forecasts is accuracy. Recall gauges the model's capacity to discover every positive

instance in the dataset, whereas precision assesses the model's accuracy in identifying positive examples. Recall and precision are balanced in the F1 score. When combined, these measures provide a thorough picture of the diagnostic performance of the model. Table 3 shows the model's evaluation metrics, where the accuracy, precision, recall, F1-score, and Cohen Kappa score are 98.36%, 98.39%, 98.37%, 98.36%, and 0.9836, respectively.

3.5.1.5 Training and Validation Metrics Visualization

When a deep learning model performs remarkably well on training data but poorly on unseen data, it is referred to as overfitting. It is crucial to track the model's performance over time (epochs) on both training and validation datasets in order to avoid this. When the accuracy and loss for both datasets are displayed side by side, it is possible to determine when the model begins to overfit. Making educated decisions like early stopping to guarantee the model stays generic and functions well on fresh, untested data is made easier with the help of such visual insights. Figure 7 shows the visualization of the model's loss and accuracy after training and validating the model. The left part of The left part of Figure 7 shows the accuracy of the model, while the right part shows the model's loss.

3.5.2 Implementation of EfficientNet Architecture

Performance accuracy and computing economy were balanced in the design of the EfficientNet architecture, namely the EfficientNetB7 variation. Its balanced scaling of model depth, width, and resolution makes this architecture stand out in particular. Grayscale mammography image processing uses this architecture. This adaptation required changing the input shape to fit the size of the mammography images (rows, columns, 1) and setting `include_top` to false to configure the network's head with the layers.

Metric	Value (%)
Accuracy	0.9836
Precision	0.9839
Recall	0.9837
F1 Score	0.9836
Cohen Kappa Score	0.9836

Table 3. Model's Evaluation Metrics for Custom Optimized CNN

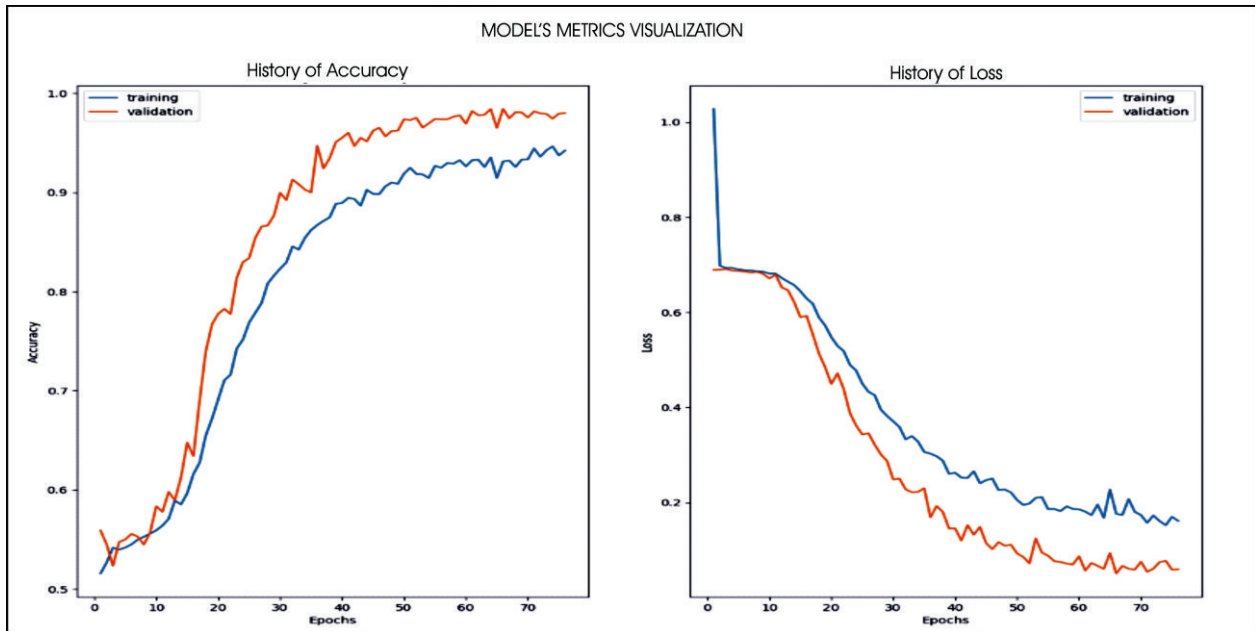


Figure 7. Model's Accuracy and Loss

An additional layer, called GlobalAveragePooling2D, was added to the model to help focus on the key features and reduce the spatial dimensions. The 'tanh' activation function was used to create a dense layer comprising 512 neurons, which introduced non-linearity and facilitated the detection of complicated patterns. Next is a dropout layer with a rate of 0.25, which is essential for regularization and avoiding overfitting. Using the sigmoid activation function, the dense layer, the last layer, had a single neuron and was perfect for binary classification tasks like differentiating between benign and malignant tumors.

3.5.2.1 Model's Summary and Training

The model was constructed, then compiled using the Adam optimizer and BinaryCrossentropy as the loss function, a popular option for binary classification problems. A clear picture of the model's complexity and structure was provided by the model summary, which included an overview of the layers, parameters, and shapes. The training process was carried out for 100 epochs with a batch size of 128 on the preprocessed dataset. This decision was predicated on finding a compromise between computing efficiency and enough learning across the training process. The inclusion

of validation splits allowed to check the model's performance on unseen data during training, guaranteeing it could identify any signs of overfitting early.

Figure 8 shows the visualization of the training process after training and validating the model. The left part of Figure 8 shows the history of accuracy of the model, while the right part shows the model's loss.

3.5.2.2 Performance Evaluation

A direct measure of the model's efficacy in diagnosing breast cancer from mammography images was provided by the loss value and accuracy obtained. After training, the model's performance was assessed on the test dataset. This assessment was essential to comprehend how well the model generalized to new data, a critical indicator of its possible utility in a real-world clinical setting.

3.5.2.3 Metrics Analysis

A number of performance indicators were calculated, as shown in Table 4, in order to fully evaluate the diagnostic capabilities of the model. Accuracy, precision, recall, F1 score, and Cohen Kappa score were among them. A unique perspective on the model's ability to accurately identify mammography pictures was provided by each statistic, which balanced the trade-offs between false

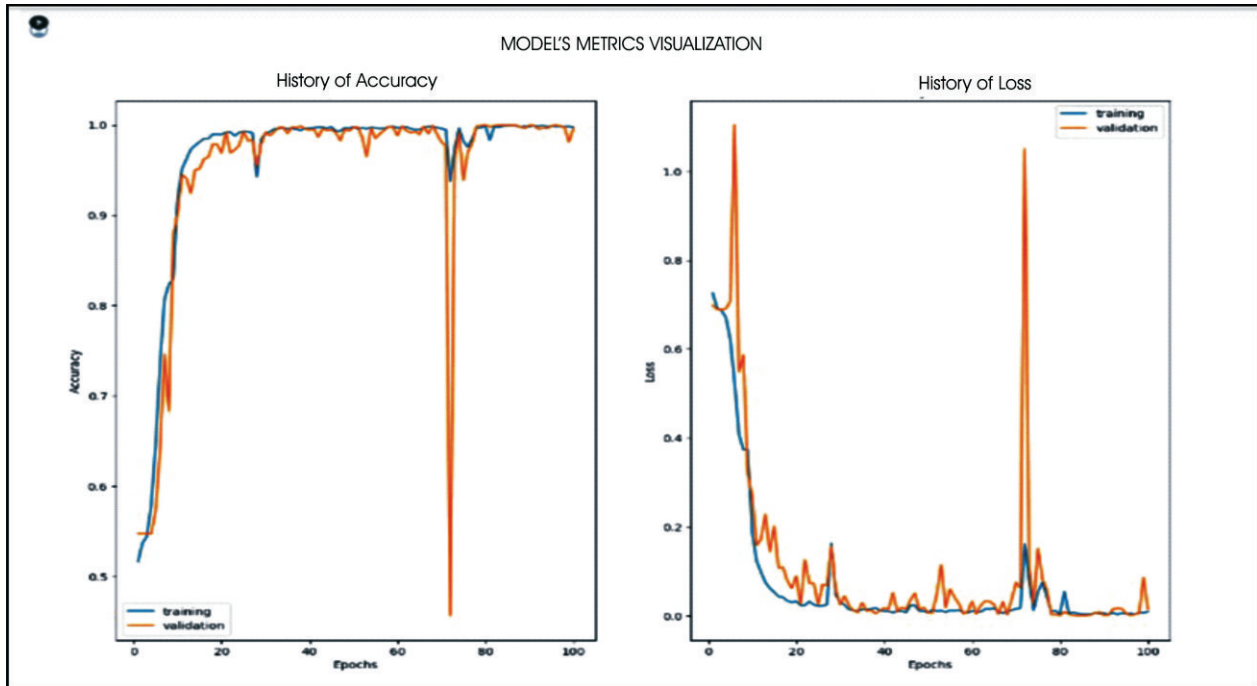


Figure 8. Training Process Visualization

Metric	Value (%)
Accuracy	0.9977
Precision	0.9914
Recall	0.9965
F1 Score	0.9939
Cohen Kappa Score	0.9954

Table 4. Classification Metrics Summary of EfficientNet

positives and negatives. A thorough categorization report also offered information on how well the model performed for each class (malignant and benign), which is essential for clinical decision-making. Table 4 shows the classification metrics summary with accuracy, precision, recall, F1-score, and Cohen kappa score of 0.9977, 0.9914, 0.9965, 0.9939, and 0.9954, respectively.

3.5.3 Implementation of ResNet Architecture

ResNet50 can learn from deeper neural networks without experiencing the vanishing gradient problem, it is a preferable architecture for neural networks. Deeper network training is made easier by ResNet50's well-known residual connections. To handle the grayscale mammography pictures in the solution, this study modified ResNet50. To achieve this, the input shape was modified to match the dimensions of the data, and the

include_top parameter was set to False.

The GlobalAveragePooling2D layer, the dropout layer with a 25% dropout rate, the dense layer with 512 neurons activated by 'tanh,' and the final dense layer with a 'sigmoid' activation function for binary output were added to the model to make it appropriate for this binary classification task. These extra layers were designed to process the ResNet50 extracted features and determine the final classification outcome.

3.5.3.1 Model Summary and Training with Early Stopping

The Binary Cross entropy loss function and the Adam optimizer were used to construct the ResNet50 model. The amount of parameters and the network's structure were explained by the model overview. An early termination mechanism was incorporated into the training program. By keeping an eye on validation accuracy and stopping training if there was no discernible progress, this method avoided overfitting and maximized the use of computational resources. Figure 9 shows the training process visualization.

3.5.3.2 Performance Evaluation

An important stage in determining the ResNet50 model's

MODEL'S METRICS VISUALIZATION

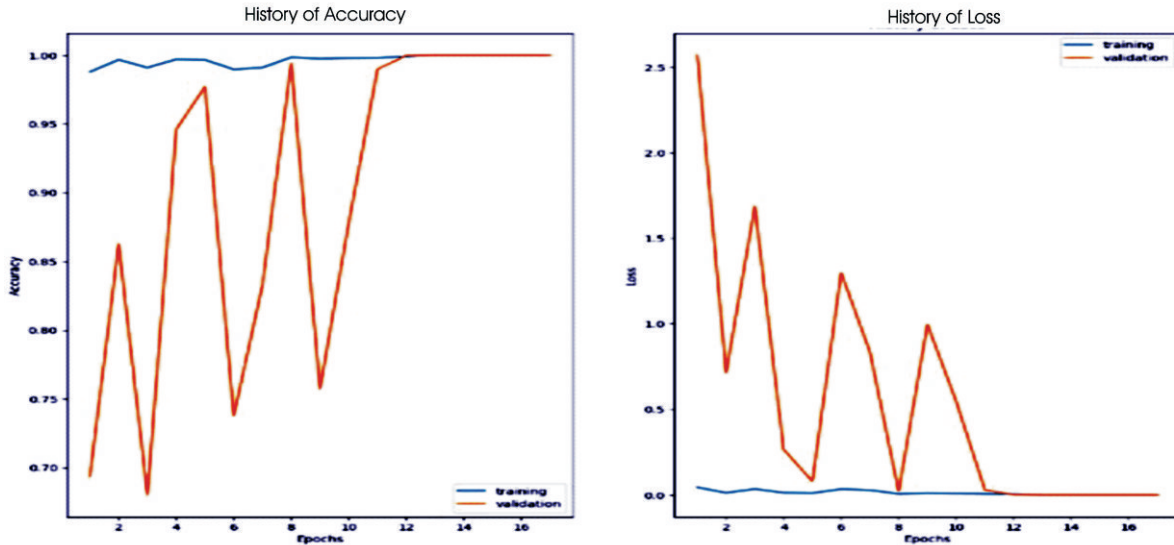


Figure 9. Training Process Visualization

efficacy was evaluating how well it performed on the test dataset. The obtained accuracy and loss values provided a clear indicator of how well the model could distinguish between benign and malignant mammogram pictures.

3.5.3.3 Metrics Analysis

A thorough analysis of the ResNet50 model was conducted utilizing a number of performance indicators. These metrics, which comprised Cohen Kappa Score, F1 score, accuracy, precision, and recall, each offered insightful information about various facets of the model's classification performance. The model's precision, recall, and F1 score for every class were further described in a classification report, providing a more in-depth understanding of the model's functionality. Table 5 shows the summary of classification metrics with accuracy, precision, recall, F1-score, and Cohen Kappa scores of 0.9987, 0.9934, 0.9967, 0.9950, and 0.9955, respectively.

3.6 Comparison of Models

The Custom Optimized CNN, EfficientNet, and ResNet architectures of the three models used to diagnose breast cancer are compared. Their classification metrics, training dynamics, and computational efficiency serve as the basis for comparison. Table 6 shows the comparison of the performances of the three CNN

Metric	Value
Accuracy	0.9987
Precision	0.9934
Recall	0.9967
F1 Score	0.9950
Cohen Kappa Score	0.9955

Table 5. Classification Metrics Summary of ResNet

	Accuracy	Precision	Recall	F1-score	Cohen Kappa score
Custom	0.9836	0.9839	0.9836	0.9837	0.9836
Optimized CNN					
EfficientNet	0.9977	0.9914	0.9965	0.9939	0.9954
ResNet	0.9987	0.9934	0.9967	0.9950	0.9955

Table 6. Comparison of Model Performance Metrics

multiple architectures.

Custom Optimized CNN: This model achieved an accuracy of 0.9836, precision of 0.9839, recall of 0.9836, F1 score of 0.9837, and Cohen Kappa score of 0.9836. Its classification report showed nearly perfect scores across both classes, with a slight variance in the benign category.

EfficientNet: The EfficientNet model exhibited almost perfect scores across all metrics, with accuracy, precision, recall, F1 score, and Cohen Kappa score of 0.9977, 0.9914, 0.9965, 0.9939, and 0.9954, respectively. Its classification report reflected these

scores, showing equally high performance in both categories.

ResNet50: ResNet50 performed best across all metrics, achieving accuracy, precision, recall, F1 score, and Cohen Kappa score of 0.9987, 0.9934, 0.9967, 0.9950, and 0.9955. The classification report confirmed its flawless performance in distinguishing between benign and malignant cases.

3.6.1 Training Dynamics and Computational Efficiency

Custom Optimized CNN: This model's training was robust, showing a gradual and consistent learning curve. It balanced computational efficiency well, making it a viable option for environments with limited computational resources.

EfficientNet: EfficientNet's training demonstrated rapid convergence to high accuracy, indicating its efficiency in learning features. The balance between model complexity and computational demands makes it an attractive choice for real-world applications.

ResNet50: ResNet50, with early stopping, showed an effective training process. The early stopping mechanism ensured that the model did not overfit and optimized the use of computational resources. Its metrics indicate that it can learn complex patterns without needing extensive training time.

3.6.2 Advantages and Implications

It is clear from a comparison of these models that each has advantages. When computing resources are limited, the Custom Optimized CNN is a great option since it provides good performance at a lower complexity.

Because EfficientNet balances computational efficiency with performance, it is appropriate for applications requiring high accuracy without using excessive amounts of resources. ResNet50 is clearly the most accurate model, with flawless scores.

The deployment environment's unique requirements and limitations will determine which of these models is selected. ResNet50 is the best option in situations where computational resources are not a constraint. EfficientNet acts as a medium ground, balancing efficiency and performance, and the Custom Optimized

CNN provides a very viable option in instances when computational resources are limited.

Conclusion

This study presents a novel method of diagnosing breast cancer by building a complex model that makes use of three CNN architectures, including ResNet, a custom optimized CNN, and EfficientNet architectures. The goal of this study is to determine the best architecture for precise breast cancer diagnosis by evaluating the performance of these several CNN architectures. The dataset used in this study was obtained from the MIAS dataset available on the Kaggle repository. With less computational complexity and 0.9932 accuracy, the Custom Optimized CNN is appropriate for environments with limited resources. With 0.9977 accuracy and good processing efficiency, EfficientNet performed admirably and is perfect for real-world applications. ResNet50.

The Custom Optimized CNN performed remarkably well, especially when taking into account its lower processing complexity. Its accuracy of 0.9836 is impressive for a model that strikes a compromise between resource efficiency and performance. EfficientNet achieved an accuracy of 0.9977, demonstrating exceptional performance. Because of its exceptional performance and balance between computational efficiency and fast performance, it is a great option for real-world applications where high precision is needed but resources are restricted. ResNet50 demonstrated its outstanding capacity to understand intricate patterns and provide precise classifications by achieving flawless scores on all measures. This results in higher computing demands.

References

[1]. Abdelli, A., Saouli, R., Djemal, K., & Youkana, I. (2020, November). Combined datasets for breast cancer grading based on multi-CNN architectures. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (pp. 1-7). IEEE.

<https://doi.org/10.1109/IPTA50016.2020.9286653>

[2]. Alanazi, S. A., Kamruzzaman, M. M., Islam Sarker, M. N., Alruwaili, M., Alhwaiti, Y., Alshammari, N., & Siddiqi, M.

H. (2021). Boosting breast cancer detection using convolutional neural network. *Journal of Healthcare Engineering*, 2021(1), 5528622.

<https://doi.org/10.1155/2021/5528622>

[3]. Altan, G. (2020). Deep learning-based mammogram classification for breast cancer. *International Journal of Intelligent Systems and Applications in Engineering*, 8(4), 171–176.

<https://doi.org/10.18201/ijisae.2020466308>

[4]. Burçak, K. C., Baykan, Ö. K., & Uğuz, H. (2021). A new deep convolutional neural network model for classifying breast cancer histopathological images and the hyperparameter optimisation of the proposed model. *The Journal of Supercomputing*, 77(1), 973-989.

<https://doi.org/10.1007/s11227-020-03321-y>

[5]. Das, H. S., Das, A., Neog, A., Mallik, S., Bora, K., & Zhao, Z. (2023). Breast cancer detection: Shallow convolutional neural network against deep convolutional neural networks based approach. *Frontiers in Genetics*, 13, 1097207.

<https://doi.org/10.3389/fgene.2022.1097207>

[6]. El Adoui, M., Mahmoudi, S. A., Larhman, M. A., & Benjelloun, M. (2019). MRI breast tumor segmentation using different encoder and decoder CNN architectures. *Computers*, 8(3), 52.

<https://doi.org/10.3390/computers8030052>

[7]. Ereken, O. F., & Tarhan, C. (2020). Breast cancer detection using convolutional neural networks. In *2022 International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT)*.

<https://doi.org/10.1109/ISMSIT56059.2022.9932694>

[8]. Houssein, E. H., Emam, M. M., & Ali, A. A. (2022). An optimized deep learning architecture for breast cancer diagnosis based on improved marine predators algorithm. *Neural Computing and Applications*, 34(20), 18015-18033.

<https://doi.org/10.1007/s00521-022-07445-5>

[9]. Khan, M. H., Boodoo-Jahangeer, N., Dullull, W., Nathire, S., Gao, X., Sinha, G. R., & Nagwanshi, K. K. (2021). Multi-class classification of breast cancer

abnormalities using Deep Convolutional Neural Network (CNN). *Plos One*, 16(8), e0256500.

<https://doi.org/10.1371/journal.pone.0256500>

[10]. Mahmood, T., Li, J., Pei, Y., Akhtar, F., Rehman, M. U., & Wasti, S. H. (2022). Breast lesions classifications of mammographic images using a deep convolutional neural network-based approach. *Plos One*, 17(1), e0263126.

<https://doi.org/10.1371/journal.pone.0263126>

[11]. Nawaz, M., Sewissy, A. A., & Soliman, T. H. A. (2018). Multi-class breast cancer classification using deep learning convolutional neural network. *International Journal of Advanced Computer Science and Applications*, 9(6), 316-332.

[12]. Niu, J., Li, H., Zhang, C., & Li, D. (2021). Multi scale attention based convolutional neural network for classification of breast masses in mammograms. *Medical Physics*, 48(7), 3878-3892.

<https://doi.org/10.1002/mp.14942>

[13]. Oyelade, O. N., & Ezugwu, A. E. (2021). A bioinspired neural architecture search based convolutional neural network for breast cancer detection using histopathology images. *Scientific Reports*, 11(1), 19940.

<https://doi.org/10.1038/s41598-021-98978-7>

[14]. Raza, A., Ullah, N., Khan, J. A., Assam, M., Guzzo, A., & Aljuaid, H. (2023). DeepBreastCancerNet: A novel deep learning model for breast cancer detection using ultrasound images. *Applied Sciences*, 13(4), 2082.

<https://doi.org/10.3390/app13042082>

[15]. Rehman, K. U., Li, J., Pei, Y., Yasin, A., Ali, S., & Saeed, Y. (2021). Architectural distortion-based digital mammograms classification using depth wise convolutional neural network. *Biology*, 11(1), 15.

<https://doi.org/10.3390/biology11010015>

[16]. Samala, R. K., Chan, H. P., Hadjiiski, L., Helvie, M. A., Richter, C. D., & Cha, K. H. (2018). Breast cancer diagnosis in digital breast tomosynthesis: Effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, 38(3), 686-696.

<https://doi.org/10.1109/TMI.2018.2870343>

[17]. Samee, N. A., Atteia, G., Meshoul, S., Al-antari, M. A., & Kadah, Y. M. (2022). Deep learning cascaded feature selection framework for breast cancer classification: Hybrid CNN with univariate-based approach. *Mathematics*, 10(19), 3631.

<https://doi.org/10.3390/math10193631>

[18]. Shivhare, E., & Saxena, V. (2021). Breast cancer diagnosis from mammographic images using optimized feature selection and neural network architecture. *International Journal of Imaging Systems and Technology*, 31(1), 253-269.

<https://doi.org/10.1002/ima.22467>

[19]. Tasnim, Z., Shamrat, F. J. M., Islam, M. S., Rahman, M. T., Aronya, B. S., Muna, J. N., & Billah, M. M. (2021). Classification of breast cancer cell images using multiple convolution neural network architectures. *International Journal of Advanced Computer Science and Applications*, 12(9), 308-315.

<https://doi.org/10.14569/IJACSA.2021.0120934>

[20]. Vijayakumar, K., Kadam, V. J., & Sharma, S. K. (2021). Breast cancer diagnosis using multiple activation

deep neural network. *Concurrent Engineering*, 29(3), 275-284.

<https://doi.org/10.1177/1063293X211025105>

[21]. Xie, L., Zhang, L., Hu, T., Huang, H., & Yi, Z. (2020). Neural networks model based on an automated multi-scale method for mammogram classification. *Knowledge-Based Systems*, 208, 106465.

<https://doi.org/10.1016/j.knosys.2020.106465>

[22]. Yektaei, H., Manthouri, M., & Farivar, F. (2019). Diagnosis of breast cancer using multiscale convolutional neural network. *Biomedical Engineering: Applications, Basis and Communications*, 31(5), 1950034.

<https://doi.org/10.4015/S1016237219500340>

[23]. Yurttakal, A. H., Erbay, H., İközceli, T., & Karaçavuş, S. (2020). Detection of breast cancer via deep convolution neural networks using MRI images. *Multimedia Tools and Applications*, 79(21), 15555-15573.

<https://doi.org/10.1007/s11042-019-7479-6>

[24]. Zhang, C., Zhao, J., Niu, J., & Li, D. (2020). New convolutional neural network model for screening and diagnosis of mammograms. *PLoS One*, 15(8), e0237674.

<https://doi.org/10.1371/journal.pone.0237674>

ABOUT THE AUTHORS

Ayobami Ekundayo is a Lecturer in the Department of Computer Science at the School of Information and Communication Technology, Federal University of Technology Minna, Niger State, Nigeria. She is currently pursuing her PhD at the university. She received both her B.Sc. and M.Sc. degrees in Computer Science from the University of Ilorin, Kwara State, Nigeria. Her research interests include Data Mining. She is a member of the Computer Professionals Registration Council of Nigeria (CPN) and has published several research papers in reputable journals and conferences.



Dr. Enesi Femi Aminu is a Senior Lecturer at the Department Computer Science, School of Information and Communication Technology, Federal University of Technology (FUT), Minna, Nigeria. He has taught so many courses, which include Artificial Intelligence, Expert Systems, Operating Systems, and Advanced Database Design. He obtained his PhD in Computer Science from FUT Minna, Nigeria. Similarly, he equally obtained both his B.Sc. and M.Sc. degrees in Computer Science from the University of Jos, Jos, and Ahmadu Bello University, Zaria, Nigeria, respectively. His research interest is on Knowledge Representations specifically Ontology Design and Semantic Search, Artificial Intelligence and Machine Learning. He is also a member of the Computer Professional Registration Council of Nigeria (CPN), the Nigeria Computer Society (NCS), and the International Association of Computer Science and Information Technology (IACSIT). He has successfully published over 35 research peer-reviewed scientific articles in different reputable publication outlets, which include journals, international conferences, and book chapters.



Winifred Omone Saliu was a Student who graduated in 2023 from the Department of Computer Science, School of Information and Communication Technology, Federal University of Technology Minna, Niger State, Nigeria. She focuses on the application of machine learning in the health sector, which served as the primary motivation. Her research interests are Machine Learning and Artificial Intelligence.



Dr. Oluwaseun Adeniyi Ojerinde is a Senior Lecturer in the Department of Computer Science, School of Information and Communication Technology, Federal University of Technology, Minna, Nigeria. He earned his B.Sc. in Computer Technology from Babcock University, Nigeria, in 2006 and his M.Sc. in Mobile Communication Systems from Loughborough University, England, in 2008. He also obtained his Ph.D. in Mobile Communication Systems from Loughborough University, England, in 2014. He is an expert in Mobile Communications and Computer Science, with extensive experience in Blockchain Technology, Digital Transformation and Innovation, Process Optimization, Antenna Systems, and Radio Propagation. His research interests include Encryption and Blockchain Technology for Data Management, Process Optimization, 5G Technology, Antenna Systems and Propagation, On-Body Systems, Multiple Input Multiple Output (MIMO) Systems, Telecommunications, Specific Absorption Rate (SAR), Radiation, and Computer Science Education. He is a member of CPN and IEEE.



Uchenna Cosmas Ugwuoke is currently pursuing his Ph.D. and working as a Lecturer I in the Department of Computer Science, Federal University of Technology, Minna, Niger State, Nigeria, where he has been employed since 2012. He obtained a Bachelor of Technology degree in Mathematics and Computer Science from the Federal University of Technology, Minna, Niger State, Nigeria, and a Master's degree in Computer Science from the University of Ibadan, Oyo State, Nigeria. He has been a member of the Nigeria Computer Society since 2011. His research areas focus on Data Analytics and Machine Learning.



IMPACT OF ARTIFICIAL INTELLIGENCE ON CYBER SHOPPING IN KANNIYAKUMARI DISTRICT

By

GOLDIE SHEILA JESOLIT C. S. *

CHITHRA JAMES **

*-** Department of Commerce, Women's Christian College, Nagercoil, Tamil Nadu, India.

<https://doi.org/10.26634/jds.2.2.21264>

Date Received: 14/10/2024

Date Revised: 28/10/2024

Date Accepted: 06/11/2024

ABSTRACT

Cyber shopping worldwide has developed significantly, especially after the COVID-19 pandemic. Due to rapid technological advances, Artificial Intelligence (AI) plays an increasingly dynamic role in cyber shopping. AI in cyber shopping tracks customers' choices, buying patterns, preferences, purchase frequency, and spending on products, helping meet the diverse needs of cyber shoppers. The study aims to examine the impact of AI on cybershoppers in Kanniyakumari District of Tamil Nadu, India. The primary data collection was through the convenience sampling method, and the sample size is 200. Secondary data collection was from journals, books, websites, and databases. Age plays a substantial role, as age-related differences explain why younger and older consumers display distinct preferences and levels of comfort with these technologies. Amazon is the most frequently used AI-based cyber shopping platform by customers, while Indiamart is the least used. The findings suggest that demographic factors are crucial in shaping customer preferences and engagement with online shopping platforms.

Keywords: Artificial Intelligence, Cyber Shopping, Shopping Platform, Technology, E-Commerce, Customer Choices, COVID-19 Pandemic, Age-Related Differences.

INTRODUCTION

Cyber shopping has grown exponentially. Continuous increases in users and high-speed internet access have significantly contributed to this growth. Customers benefit from various advantages, like convenience, lower prices, time savings, and instant access to all necessary information. However, a few issues persist, such as inadequate customer service, insufficient return policies, and safety concerns (Jangra & Jangra, 2022). Artificial Intelligence (AI) enables non-human machines or artificial entities to perform tasks, solve problems, communicate, interact, and behave logically, similar to biological humans (Zúñiga et al., 2023). AI plays a vital

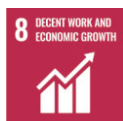
role in shaping the buying patterns of cyber shoppers worldwide. It transforms cyber shopping by interpreting external data accurately, learning from this data, and using those insights to achieve specific goals and tasks through flexible adaptation. E-commerce uses AI techniques, systems, tools, or algorithms to support activities related to buying and selling products or services online (Bawack et al., 2022). Therefore, with AI, customers purchase from cyber shopping platforms more efficiently. AI also focuses on the development of the advantages of cyber shopping and turning the disadvantages of cyber shopping into advantages.

1. Review of Literature

Lari et al. (2022) identified some key applications of AI in e-commerce by reviewing research articles from various sources. The study concluded that AI holds a powerful impact on e-commerce and aims to bring new trends in the near future. AI can change the companies and the



This paper has objectives related to SDGs



framework within which the e-commerce websites work and provide a new touch, thereby replacing the traditional and physical way of shopping for their customers.

Pillarisetty and Mishra's (2022) outlined the different AI-based technologies that impact online fashion retail. It is found that AI can help in improving several service features, such as tailoring fits with mix and match, 24 hours of customer assistance, and personalization services with chatbots. This improved efficiency reduces returns and improves repeat purchases. Personalized recommendations are known to positively influence customer loyalty. 'Behavioural Analytics' can be a tool to understand the behavior of online customers and are used to model customers's behavior before purchase. This will save company time and money, and one should understand this process in order to improve personalization and the overall customer experience.

Singh et al. (2023) focused on C2C business, in which customers or ultimate users buy the goods and services from online retailers or marketplaces. Simulated intelligence innovation is one thing that is turning out to be more adaptable and versatile to human idea designs. In the e-retail domain, AI has shown a greater scope to retailers for encouraging more participation in the e-commerce business. With the assistance of AI, retail organizations can make their products more visual to their segmented customers as they realize what else they can look for on their commercial platforms.

Subbaiah et al. (2024) investigated the factors that influence artificial intelligence's practical implacability to better understand how it affects consumers' online purchase plans. The findings of the study provided support for the hypothesis that changing customers' online shopping behaviors necessitates a comprehensive understanding of the factors that are important to them. AI tools appear to be an effective tool for accomplishing the goal because of the valuable data that they provided on a variety of variables that influence the behavioral intentions of customers. Marketing can reap benefits from facilitating conditions such as longer purchase durations and increased dependence on

social media.

Kumar and Senthilmurugan (2024) evaluated the awareness level of consumers about AI and its impact on purchase decisions based on demographic characteristics. The study revealed that there is no significant association between gender and awareness about AI's impact on online shopping. However, there is a significant association between educational level and the time spent on the internet with the influence of AI on online shopping.

2. Objectives

- To examine the reasons of customers to purchase through cyber shopping with the aid of AI.
- To analyze the impact AI has created on cyber shopping among customers.

3. Hypothesis

- *H01*: There is no significant difference between the customer's age and the factors involved in the usage of AI in cyber shopping.
- *H02*: There is no significant difference among the mean ranks towards the AI-based cyber websites used by the customers.
- *H03*: There is no significant difference between customers' gender and the challenges faced in the usage of AI in cyber shopping.

4. Research Methodology

This study opts for a convenience sampling method to collect the data for the research among 200 individuals in Kanniyakumari District, Tamil Nadu, India. Primary data and secondary data collection are the tools used for data collection. The collection of primary data is through a well-structured questionnaire through Google Forms. In secondary data collection, journals, books, databases, and transcripts were the sources of information. The statistical tools used for the research are percentage analysis, one-way ANOVA, the Friedman test, and an independent sample t-test with the aid of SPSS software.

5. Limitations of the Study

5.1 Sample Limitation

This study uses a convenience sampling method, which

may not accurately represent the entire population of Kanniyakumari District, Tamil Nadu, India. This can limit the generalizability of the findings to a broader audience.

5.2 Data Collection Constraints

Since primary data collection is done through Google Forms, respondents who are less familiar with digital platforms or who lack access to the internet might be underrepresented, potentially leading to sample bias.

5.3 Limited Scope of Factors and Variables

This study specifically examines age and gender as demographic factors, but it may not account for other important variables that could influence customer interaction with AI in cyber shopping.

5.4 Rapidly Evolving Technology

AI and cyber shopping technologies evolve quickly, so findings from this study may become outdated as new advancements and trends emerge in the e-commerce sector.

6. Analysis and Interpretation

6.1 Demographic Profile

Table 1 shows the demographic profile of the population belonging to Kanniyakumari District, Tamil Nadu, India.

Table 1 shows that 29.5 percent (59) of the respondents belong to the age group of 34-41. It is followed by the respondents belonging to the age groups of 42-49, 26-33, and 50 and above with 28.0 percent (56), 19.5 percent (39) and 15.0 percent (30), respectively. Only 8.0 percent (16) of the respondents belong to the age group of 18-25. 57.5 percent (115) of the respondents are male, while 42.5 percent (85) of the respondents are female. The majority of the respondents, 37.0 percent (74) of the respondents, are professionals. It is followed by the respondents having an educational qualification in post-graduation, high school, and undergraduation with 20.0 percent (40), 20.0 percent (40), and 13.5 percent (27), respectively. Only 9.5 percent (19) of the respondents have completed an educational qualification in diploma programs. Most of the respondents' annual family income is between Rs. 9,00,001 and Rs. 12,00,000, with 32.0 percent (64). It is followed by those having an annual

	Particulars	Frequency	Percent
Age	18-25	16	8.0
	26-33	39	19.5
	34-41	59	29.5
	42-49	56	28.0
	50 and Above	30	15.0
	Total	200	100.0
Gender	Male	115	57.5
	Female	85	42.5
	Total	200	100.0
Educational Level	High School	40	20.0
	Under-Graduate	27	13.5
	Post-Graduate	40	20.0
	Professional Courses	74	37.0
	Diploma Programs	19	9.5
	Total	200	100.0
Annual Income	Rs. 3,00,000 and Below	43	21.5
	Rs. 3,00,001 to 6,00,000	27	13.5
	Rs. 6,00,001 to 9,00,000	38	19.0
	Rs. 9,00,001 to 12,00,000	64	32.0
	Above Rs. 12,00,000	28	14.0
	Total	200	100.0
Marital Status	Single	62	31.0
	Married	115	57.5
	Divorced	16	8.0
	Widow	7	3.5
	Total	200	100.0
Nature of Family	Nuclear	136	68.0
	Joint	64	32.0
	Total	200	100.0
Number of Member	Less than 4	108	54.0
	4-8	66	33.0
	More than 8	26	13.0
	Total	200	100.0
Current Accommodation Status	Rented House	77	38.5
	Own House	117	58.5
	Leased House	6	3.0
	Total	200	100.0
Taluk	Agastheeswaram	86	43.0
	Thovalai	30	15.0
	Vilavancode	42	21.0
	Thiruvattar	20	10.0
	Kalkulam	12	6.0
	Killiyoor	10	5.0
Total	200	100.0	

Source: Primary Data

Table 1. Demographic Profile

family income between Rs. 3,00,000 and below, Rs. 6,00,001 to 9,00,000, and above Rs. 12,00,000 with 21.5 percent (43), 19.0 percent (38) and 14.0 percent (28), respectively. The least number of respondents earn an annual family income of between Rs. 3,00,001 and 6,00,000, with 13.5 percent (27). 57.5 percent (115) of the respondents are married, while 31.0 percent (62) of the respondents are unmarried. 8.0 percent (16) of the respondents are divorced, and another 3.5 percent (7) of the respondents are widows. The majority of the

respondents, 68.0 (136), are from nuclear families, while 32.0 percent (64) of the respondents are from joint families. Most respondents, 54.0 percent (108), have less than four members in their family. 33.0 percent (66) of the respondents have four to eight members in their family, and only 13.0 percent (26) of the respondents have more than eight members in their family. Around 58.5 percent (117) are living in their own house. 38.5 percent (77) of the respondents are living in a rented house, and only 3.0 percent (6) of the respondents reside in a leased house. The majority, 43.0 percent (86), of the respondents belong to Agastheeswaram Taluk. The respondents belonging to Vilavancode Taluk, Thovalai Taluk, Thiruvattar Taluk, and Kalkulam Taluk are 21.0 percent (42), 15.0 percent (30), 10.0 percent (20), and 6.0 percent (12). The least number of respondents, 5.0 percent (10), belongs to Killiyoor.

6.2 Factors Involved in the Usage of AI in Cyber Shopping

Several factors influence the effective usage of AI in cyber

shopping, shaping both the customer experience and business outcomes. Table 2 shows the factors involved in the usage of AI in cyber shopping with the use of one-way ANOVA.

Table 2 shows that the p-value is less than 0.01 for the factors 'Personalization and Recommendations, Chatbots and Virtual Assistants, Predictive Analytics, Visual Search and Image Recognition', 'Fraud Detection and Security', 'Voice Commerce', 'Sentiment Analysis and Customer Feedback', 'Augmented Reality (AR) and Virtual Reality (VR)', 'Automation of Order Fulfillment', and 'Ethical Considerations and Transparency' at one percent significance level. Hence, the null hypothesis is rejected, and the alternate hypothesis of 'There is a high significant difference between the customer's age and the factors involved in the usage of AI in cyber shopping' is accepted at a 99 percent confidence level.

The significant relationship between customer age and

Factors		Sum of Squares	df	Mean Square	F	Sig.
Personalization and Recommendations	Between Groups	182.913	4	45.728	29.776	<0.001**
	Within Groups	299.467	195	1.536		
	Total	482.380	199			
Chatbots and Virtual Assistants	Between Groups	182.913	4	45.728	29.776	<0.001**
	Within Groups	299.467	195	1.536		
	Total	482.380	199			
Predictive Analytics	Between Groups	165.659	4	41.415	26.424	<0.001**
	Within Groups	305.621	195	1.567		
	Total	471.280	199			
Visual Search and Image Recognition	Between Groups	174.100	4	43.525	28.313	<0.001**
	Within Groups	299.775	195	1.537		
	Total	473.875	199			
Fraud Detection and Security	Between Groups	198.924	4	49.731	35.536	<0.001**
	Within Groups	272.896	195	1.399		
	Total	471.820	199			
Voice Commerce	Between Groups	183.550	4	45.888	29.268	<0.001**
	Within Groups	305.730	195	1.568		
	Total	489.280	199			
Sentiment Analysis and Customer Feedback	Between Groups	173.413	4	43.353	26.764	<0.001**
	Within Groups	315.867	195	1.620		
	Total	489.280	199			
Augmented Reality (AR) and Virtual Reality (VR)	Between Groups	181.386	4	45.347	29.625	<0.001**
	Within Groups	298.489	195	1.531		
	Total	479.875	199			
Automation of Order Fulfillment	Between Groups	175.604	4	43.901	26.923	<0.001**
	Within Groups	317.976	195	1.631		
	Total	493.580	199			
Ethical Considerations and Transparency	Between Groups	200.755	4	50.189	33.535	<0.001**
	Within Groups	291.840	195	1.497		
	Total	492.595	199			

Source: Primary Data

** denotes significant at 1%

Table 2. Factors Involved in the Usage of AI in Cyber Shopping

various AI-driven factors in cyber shopping, indicated by a p-value less than 0.01, suggests that age plays a crucial role in how consumers interact with and respond to AI technologies in e-commerce. Younger consumers, more comfortable with digital environments, may be more inclined to use features such as personalization and recommendations, chatbots, visual search, and voice commerce. They are typically more familiar with technology and may find these features intuitive and convenient. Conversely, older customers might prioritize security and privacy, explaining their interest in fraud detection and ethical considerations in AI. Additionally, factors like augmented and virtual reality (AR/VR) may appeal more to younger users who are accustomed to immersive and interactive experiences through gaming or social media. This age-related variation in preferences highlights how AI features in cyber shopping meet different needs across demographics, making age a significant factor in customer engagement with these technologies.

6.3 AI Based Cyber Websites Used by the Customers

AI-based cyber shopping websites use artificial intelligence to enhance user experiences, streamline shopping processes, and provide personalized services. The most commonly used AI-based cyber websites by the customers are analyzed with the aid of the Friedman test.

Table 3 shows that Amazon is the most used AI-based cyber website by the customer and hence ranked in the first position with a mean score of 3.91. This is followed by Flipkart, Myntra, Shopper's Stop, eBay, Lenskart, and Caratlane in the second, third, fourth, fifth, sixth, and seventh positions with a mean score of 3.80, 3.69, 3.67, 3.56, 3.50, and 3.47, respectively. The least used AI-based cyber website by the customers is Indiamart, with a mean score of 2.84 and hence, ranked in the eighth position.

Amazon ranks highest, with a mean score of 3.91, likely due to its advanced AI capabilities, personalized recommendations, and seamless user interface, which create a highly engaging shopping experience. Lower-ranked platforms like Indiamart, with a mean score of

Cyber Websites	N	Mean	Rank
Amazon	200	3.91	I
eBay	200	3.56	V
Myntra	200	3.69	III
Shopper's Stop	200	3.67	IV
Caratlane	200	3.47	VII
Flipkart	200	3.80	II
Indiamart	200	2.84	VIII
Lenskart	200	3.50	VI
N		200	
Chi-Square		199.769	
df		7	
Asymp. Sig.		<0.001**	

Source: Primary Data
 ** denotes significant at 1%

Table 3. AI Based Cyber Websites Used by the Customers

2.84, tend to cater more to B2B markets or niche products, which may not align with the general consumer preferences observed in this study.

The p-value is less than 0.01 at the one percent significance level, and therefore, the null hypothesis is rejected, and the alternate hypothesis, 'There is a high significant difference among mean ranks towards the AI-based cyber websites used by the customers', is accepted at the 99 percent confidence level.

The significant difference in the usage of AI-based cyber shopping websites reflects varying levels of customer preference and satisfaction with each platform's AI-driven features and user experience.

The p-value being less than 0.01 at the 1% significance level confirms that these observed differences are statistically significant, highlighting that consumers gravitate towards platforms that effectively integrate AI to enhance convenience, personalization, and security. This variation underscores how AI features play a key role in shaping user engagement across different cyber shopping websites.

6.4 Challenges Faced in The Usage of AI in Cyber Shopping

Challenges include the difficulties and obstacles encountered when implementing and using AI technologies in the online shopping environment. Table 4 shows such an issue with the aid of an independent t-test by comparing it with the independent variable gender.

Table 4 shows that the p-values for the variables 'Data

Privacy and Security Concerns', 'Bias and Discrimination in AI Algorithms', 'High Implementation and Maintenance Costs', 'Lack of Human Touch in Customer Service', 'Over-reliance on Data Quality', 'Cybersecurity Risks', 'Difficulty in Achieving Personalization Without Intrusiveness', and 'Cultural and Regional Differences' are less than 0.01 at the one percent significance level. Therefore, the null hypothesis is rejected, and the alternate hypothesis of 'There is a high significant difference between customers' gender and the challenges faced in the usage of AI in cyber shopping' is accepted at a 99 percent confidence level. This may be due to varying gender perspectives on privacy, customer interaction preferences, and security. Studies suggested women may be more concerned about data privacy and security, which could account for significant differences in perceptions related to these variables.

The p-value for the variables 'Complexity of AI' and 'Limited Creativity in AI' is more than 0.05 at the five percent significance level. Hence, the null hypothesis, 'There is no significant difference between customers' gender and the challenges faced in the usage of AI in cyber shopping', is accepted.

These aspects might be viewed as more general, technical challenges of AI that both genders experience

similarly. This suggests that while certain challenges in AI usage are influenced by gender-specific concerns, others are universally perceived across customers, regardless of gender.

7. Findings

The findings from this study reveal significant relationships between various demographic factors and the usage of AI in cyber shopping. Table 2 shows that customer age is a crucial determinant of engagement with AI technologies, as evidenced by a p-value less than 0.01 for several AI-driven features, suggesting younger consumers are more inclined to utilize tools like personalization, chatbots, and visual search, while older customers prioritize security and ethical considerations. Table 3 shows Amazon as the most utilized AI-based cyber shopping platform, with a mean score of 3.91, attributed to its robust AI capabilities and user-friendly interface, whereas Indiamart ranks lowest at 2.84, reflecting its B2B focus. The statistical significance ($p < 0.01$) of the differences among mean ranks reinforces the notion that consumer preferences for these platforms vary significantly based on their perceived effectiveness of AI integration. Finally, Table 4 shows notable gender differences in perceptions of challenges faced when using AI in cyber shopping, particularly regarding data

Challenges	Gender	N	Mean	Std. Deviation	t	Sig (Two-Tailed)
Data Privacy and Security Concerns	Male	115	2.67	1.64	10.05	0.001**
	Female	85	3.36	1.31		
Bias and Discrimination in AI Algorithms	Male	115	2.63	1.67	12.39	0.001**
	Female	85	3.36	1.31		
High Implementation and Maintenance Costs	Male	115	2.63	1.67	12.39	0.001**
	Female	85	3.36	1.31		
Lack of Human Touch in Customer Service	Male	115	2.60	1.64	9.63	<0.001**
	Female	85	3.36	1.31		
Over-reliance on Data Quality	Male	115	2.54	1.67	12.07	<0.001**
	Female	85	3.36	1.31		
Cybersecurity Risks	Male	115	2.65	1.69	13.70	0.001**
	Female	85	3.36	1.31		
Difficulty in Achieving Personalization Without Intrusiveness	Male	115	4.31	1.17	2.98	<0.001**
	Female	85	3.36	1.31		
Complexity of AI	Male	115	3.71	1.41	7.62	0.073
	Female	85	3.36	1.31		
Cultural and Regional Differences	Male	115	3.93	1.20	10.05	0.002**
	Female	85	3.36	1.31		
Limited Creativity in AI	Male	115	3.59	1.12	12.39	0.190
	Female	85	3.36	1.31		

Source: Primary data

** denotes significant at 1%,

Table 4. Challenges Faced in the Usage of AI in Cyber Shopping

privacy and security concerns, with p-values less than 0.01. In contrast, the variables "complexity of AI" and "limited creativity in AI" do not show significant differences across genders, suggesting that while some challenges are influenced by gender perspectives, others are experienced uniformly by all users. These findings underscore the importance of demographic factors in shaping user engagement and experiences with AI technologies in the e-commerce landscape.

Conclusion

In conclusion, this study highlights the significant impact of demographic factors, particularly age and gender, on consumer interactions with AI technologies in cyber shopping. This study reveals that younger consumers are more adept at utilizing AI-driven features such as personalization, chatbots, and visual search, which enhance their shopping experience, while older consumers exhibit greater concern for security and ethical implications. The prominence of Amazon as the leading AI-based cyber shopping platform underscores the importance of effective AI integration in driving customer preference and satisfaction. Furthermore, this study identifies notable gender differences in perceptions of challenges related to AI usage, particularly regarding data privacy and security, while also indicating that certain technical challenges are universally experienced. These findings emphasize the necessity for e-commerce platforms to tailor their AI features to meet diverse consumer needs and preferences, ultimately enhancing user engagement and fostering a more inclusive and secure shopping environment. As the landscape of cyber shopping continues to evolve, understanding these demographic influences will be crucial for businesses aiming to leverage AI technologies effectively and responsively. The future scope of this study includes expanding the research to a larger, more diverse population across different geographic regions to gain insights into varying consumer attitudes toward AI in cyber shopping. Longitudinal studies could explore how these attitudes change over time with technological advancements. Future research could also address the ethical implications of AI, focusing on consumer trust and

algorithm transparency.

References

- [1]. Jangra, G., & Jangra, M. (2022, September). Role of Artificial Intelligence in online shopping and its impact on consumer purchasing behaviour and decision. In *2022 Second International Conference on Computer Science, Engineering and Applications (ICCSEA)* (pp. 1-7). IEEE.
<https://doi.org/10.1109/ICCSEA54677.2022.9936374>
- [2]. Gil de Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A scholarly definition of artificial intelligence (AI): Advancing AI as a conceptual framework in communication research. *Political Communication*, 41(2), 317-334.
<https://doi.org/10.1080/10584609.2023.2290497>
- [3]. Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022). Artificial Intelligence in E-Commerce: A bibliometric study and literature review. *Electronic Markets*, 32(1), 297-338.
<https://doi.org/10.1007/s12525-022-00537-z>
- [4]. Lari, H. A., Vaishnava, K., & Manu, K. S. (2022). Artificial Intelligence in E-commerce: Applications, implications and challenges. *Asian Journal of Management*, 13(3), 235-244.
<http://doi.org/10.52711/2321-5763.2022.00041>
- [5]. Pillarisetty, R., & Mishra, P. (2022). A review of AI (Artificial Intelligence) tools and customer experience in online fashion retail. *International Journal of E-Business Research (IJEER)*, 18(2), 1-12.
<https://doi.org/10.4018/IJEER.294111>
- [6]. Singh, N., Chaturvedi, D. D., Mittal, A., & Mittal, A. (2023). Impact of Artificial Intelligence in online customer satisfaction: An empirical study using multiple regression analysis. *European Economic Letters*, 13(1), 396-409.
- [7]. Subbaiah, P. V., Jyothsna, M., Manjushree, p., & Kondamudi, S. G. (2024). Exploring the influence of Artificial Intelligence (AI) on online purchase decisions: In case of consumer's prospective. *International Journal of Intelligent Systems and Applications in Engineering*, 12(10), 13-20.

[8]. Kumar, C. M., & Senthilmurugan, T. (2024). Impact on Intelligence. *Educational Administration: Theory and Practice*, 30(5), 2292-2295.

ABOUT THE AUTHORS

Goldie Sheila Jesolif C. S. has been a Research Scholar at Women's Christian College, Nagercoil (affiliated with Manonmaniam Sundaranar University), Tamil Nadu, India, since October 2021. She has completed her Master's degrees in Commerce and Business Administration and has also passed the UGC-NET examination conducted by the NTA in the subjects of Commerce and Management. She has published numerous research papers and received the "Best Paper" award at a national conference held by the SRM Institute of Science and Technology, Chennai, Tamil Nadu, India. Additionally, she has attended various conferences and seminars.



Dr. Chithra James has been a Professor at Women's Christian College, Nagercoil (affiliated with Manonmaniam Sundaranar University), Tamil Nadu, India, since 1992. Her areas of interest include Income Tax, Accountancy, and Computer-Related Subjects. She has a specialized interest in conducting research on Consumer Behavior and has been guiding Ph.D. students since 2015. She has published numerous research papers in national and international journals and has attended many conferences and seminars.



AN INTERACTIVE VISUALIZATION AND DATA ANALYSIS OF SUPERMARKET STORE

By

LIKITA H. L. *

HIRESH YADAV **

ANJALI HEDAHO ***

SWAGOTA BERA ****

SIDDHARTHA CHOUBEY *****

*-***** Shri Shankaracharya Engineering College, Bhilai, Chhattisgarh, India.

<https://doi.org/10.26634/jds.2.2.20814>

Date Received: 16/05/2024

Date Revised: 05/09/2024

Date Accepted: 04/11/2024

ABSTRACT

In the ever-evolving landscape of retail, the utilization of data-driven insights plays a pivotal role in enhancing decision-making processes. This paper presents an advanced approach to the development and implementation of an interactive visualization system tailored specifically for the analysis of data within a supermarket store environment. In order to predict the sales of a business, an intelligent model was built using Linear Regression, LASSO Regression, and XGBoost techniques, which have been shown to be more effective than existing models. The proposed system integrates data pre-processing, feature engineering, and algorithmic enhancements to analyze customer behavior, sales trends, and inventory efficiency. Key performance metrics like RMSE and R-squared validate the efficacy of the models, highlighting the XGBoost algorithm's exceptional performance. This study aims to optimize decision-making in inventory management, marketing strategies, and sales forecasting, ensuring data-driven insights for improved supermarket operations.

Keywords: Interactive Visualization, Data Analysis, XG-Boost Regression, Linear-Regression, LASSO-Regression, Random Forest, Sales, Prediction, Supermarket.

INTRODUCTION

The world of supermarkets thrives on understanding its customers and products. In this data-driven era, supermarket chains have access to a wealth of information about their operations. The Python programming language was used for this application, along with Jupyter Notebook as a development tool. Machine learning components, such as supervised learning tasks, were employed in the creation of the application, with regression tasks being a key focus. The primary objective was to forecast future corporate retail

sales. The process involved the collection of raw data, feature engineering, data preprocessing, model prototyping, and evaluation of techniques.

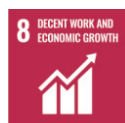
The knowledge gained through supervision helps in understanding the flow of data, including sales prices and related trends. The regression analysis relies on a variety of techniques to forecast prices. It is used to perform tasks such as data cleaning, data transformation, and implementing XGBoost algorithms. This study used XGBoost to develop a prediction model, which was tested on the database to predict the sales of products from an outlet (Suma & Hills, 2020; Wang, 2019).

1. Objectives of the Study

- To analyze sales trends and patterns to identify popular products and customer preferences.
- To assess inventory management efficiency and



This paper has objectives related to SDGs



identify areas for improvement.

- To explore customer segmentation based on purchasing behavior and demographics.
- To visualize geographical distribution of sales and identify areas of high demand.
- To evaluate the effectiveness of marketing strategies through data-driven insights.
- To identify outliers and anomalies in sales data for further investigation.
- To enhance decision-making processes for inventory stocking and product placement.
- To communicate findings effectively through data visualization techniques for stakeholders' understanding and strategic planning.

2. Related Works

In order to predict illegal acts, sales, sanitary conditions, and home values and to assess cardiorespiratory risk using XGBoost, this study used several regression models (Nunnari & Nunnari, 2017). These items are sold at supermarket shops, and as their production is in large quantities and growing regions are more difficult to predict manually, it becomes more and more challenging (Abhinaya et al., 2023). In this case, the programming language Python is used with Julia's notebook as a tool. The features of machine learning (supervised) and regression and learning functions are used in this application (Tsoumakas, 2019). This application is mainly used to anticipate the supermarket's future sales and to understand how to increase sales and regression function forecasts using a variety of algorithms. The identification, cleaning, and transformation of the

data require effort. The profit margin of the enterprise is closely tied to accurate sales forecasts, and supermarkets are looking for an efficient forecasting method that ensures they do not lose money. This method has been shown to be more accurate than alternative methods such as decision trees and local collection.

3. Model's Architecture

Figure 1 shows the architecture of the ML model.

4. Dataset

This dataset was collected using the website kaggle.com from the internet. The dataset consists of datasets for testing and training. The test data consists of 900 data rows, and the train data consists of 100 data rows. Table 1 shows the data information. Figure 2 shows the training data. Figure 3 shows the testing data.

4.1 Pre-Processing of Data

- *Data Collection*: Relevant sources of data, including sales records, inventory, human ecology, and external factors, are converged.
- *Cleaning*: Missing values are handled, duplicates are removed, and outliers are detected and adjusted.
- *Normalization or Scaling*: Numerical features are scaled to a similar range.
- *Data Splitting*: The data is split into training, validation, and test sets.
- *Transformation*: Skewness is addressed, date/time features are handled, and necessary transformations are applied.
- *Handling Imbalanced Data*: Techniques to address class imbalance are implemented.

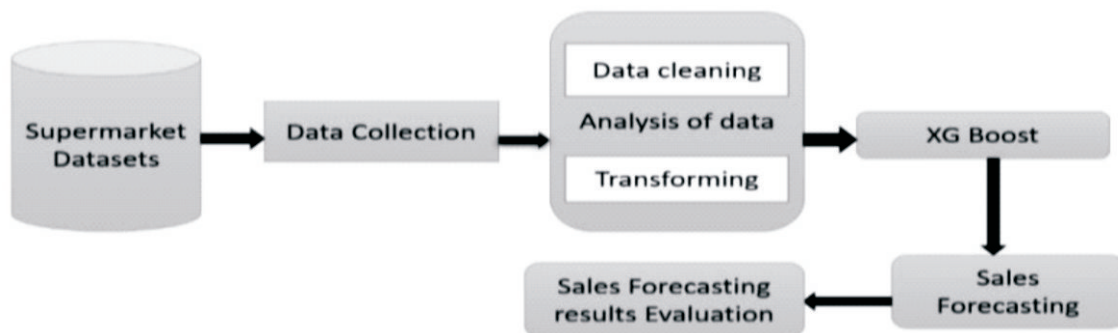


Figure 1. Architecture of ML-Model

Variable Name	Description	Datatype
Invoice ID	Unique ID of Every Customer	object
Branch	Different Branches of Supermarket Store	object
City	Different City Names From Where The Customer Belongs	object
Customer Type	Different Customer Types	object
Gender	Customer's Gender	object
Product Line	Different Category of Product	object
Unit Price	Unit Price Per Item	float64
Quantity	Number of Items Being Sold	Int64
Tax 5%	Tax on Each Item	float64
Total	Total Amount of Product That The Customer Purchase	float64
Date	Date of Purchase	object
Time	Time of Purchase	object
Payment	Different Type of Payment Methods	object
Cogs	Price of Goods That are Sold	float64
Gross margin percentage	Total Margin Percentage on Each Item	float64
Gross income	The Amount That a Business Makes Before Deductions Are Added In	float64
Rating	Customer Rating Given For Each Product Line	float64

Table 1. Data Information

	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Payment	cogs	Rating	date	day	month	year	Hour	Female	Ma	
785	0	2	1	1	0	74.58	7	26.1030	548.1630	53	1	522.06	9.0	34	4	2	2019	16	0		
873	0	2	0	1	5	89.06	6	26.7180	561.0780	9	0	534.36	9.9	17	18	1	2019	17	0		
65	0	2	0	1	3	15.87	10	7.9350	166.6350	63	0	158.70	5.8	71	13	3	2019	16	0		
902	0	2	0	0	3	27.73	5	6.9325	145.5825	77	1	138.65	4.2	84	26	3	2019	20	1		
317	2	1	0	1	0	74.79	5	18.6975	392.6475	1	0	373.95	4.9	9	10	1	2019	11	0		
...
835	0	2	1	1	1	52.38	1	2.6190	54.9990	77	0	52.38	5.8	84	26	3	2019	19	0		
192	2	1	1	0	2	87.80	9	39.5100	829.7100	66	0	790.20	9.2	74	16	3	2019	19	1		
629	0	2	1	0	1	12.09	1	0.6045	12.6945	18	1	12.09	8.2	25	26	1	2019	18	1		
559	0	2	0	0	4	72.42	3	10.8630	228.1230	80	2	217.26	8.2	87	29	3	2019	16	1		
684	1	0	0	0	5	23.08	6	6.9240	145.4040	16	2	138.48	4.9	23	24	1	2019	19	1		

900 rows × 20 columns

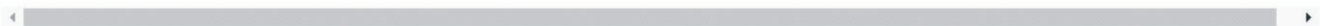


Figure 2. Training Data

	Branch	City	Customer type	Gender	Product line	Unit price	Quantity	Tax 5%	Total	Date	Payment	cogs	Rating	date	day	month	year	Hour	Female	Ma	
993	1	0	1	1	1	17.49	10	8.7450	183.6450	45	2	174.90	6.6	52	22	2	2019	18	0		
859	0	2	0	0	2	47.63	9	21.4335	450.1035	15	0	428.67	5.0	22	23	1	2019	12	1		
298	0	2	0	1	4	60.01	4	12.0020	252.0420	17	0	240.04	4.5	24	25	1	2019	15	0		
553	2	1	1	1	0	22.21	6	6.6630	139.9230	86	1	133.26	8.6	65	7	3	2019	10	0		
672	1	0	1	0	3	73.41	3	11.0115	231.2415	70	2	220.23	4.0	60	2	3	2019	13	1		
...
485	1	0	0	0	5	49.49	4	9.8980	207.8580	72	2	197.96	6.6	79	21	3	2019	15	1		
568	1	0	1	0	1	79.86	7	27.9510	586.9710	1	1	559.02	5.5	9	10	1	2019	10	1		
108	2	1	1	0	2	33.98	9	15.2910	321.1110	75	0	305.82	4.2	82	24	3	2019	10	1		
367	0	2	0	1	4	65.94	4	13.1880	276.9480	75	0	263.76	6.0	82	24	3	2019	10	0		
644	2	1	0	1	0	12.05	5	3.0125	63.2625	38	2	60.25	5.5	46	16	2	2019	15	0		

100 rows × 20 columns

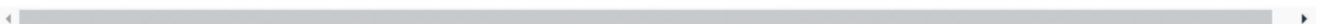


Figure 3. Testing Data

- *Visualization:* Pre-processed data is visualized to understand relationships and patterns.
- *Final Data Preparation:* The data is formatted for machine learning model input (Raizada & Saini, 2021).

4.2 Feature Engineering

In this step of feature engineering, redundant columns, such as Invoice ID and Time, were removed. To convert categorical variables to numeric representations, the label encoding technique was applied. Finally, the data was split into training and testing datasets, and the features were normalized (Niu, 2020). Figure 4 shows the bar plot between the product line and gross income.

4.3 Prototyping of Model

Regression models, including XGBoost, were trained for modeling. The XGBoost hyperparameters were tuned using grid search (Silva & Cardoso, 2005). The model's performance was analyzed using RMSE and R-squared metrics.

5. Methodology

5.1 Algorithms

5.1.1 XGBoost Algorithm

XGBoost (eXtreme Gradient Boosting) is an efficient and

dominant machine-learning algorithm renowned for its performance in supervised learning tasks, particularly in structured or tabular data and regression or classification problems. It is related to the group of ensemble learning techniques, specifically gradient boosting algorithms, which sequentially build a series of weak learners (typically decision trees) and combine their predictions to form a robust final model. Its key features include a highly optimized implementation, support for parallel processing, tree pruning techniques, and the ability to handle missing data natively. This study aims at developing a robust and precise model for the forecasting of sales in the supermarket store (Kuo et al., 2009). As in any merchandise sector, food retailers face several issues affecting sales, such as seasonality, geographical location, and marketing strategies. This study is developing a predictive model for predicting sales of products within supermarkets soon (Cumby et al., 2004; Silva & Cardoso, 2005). The ability of retailers to make decisions on the basis of data, effectively manage their stocks, and improve total performance could be improved through this model. The results of this study have a profound impact on retailers, but it also represents an impressive specimen of the disruption latent of advanced

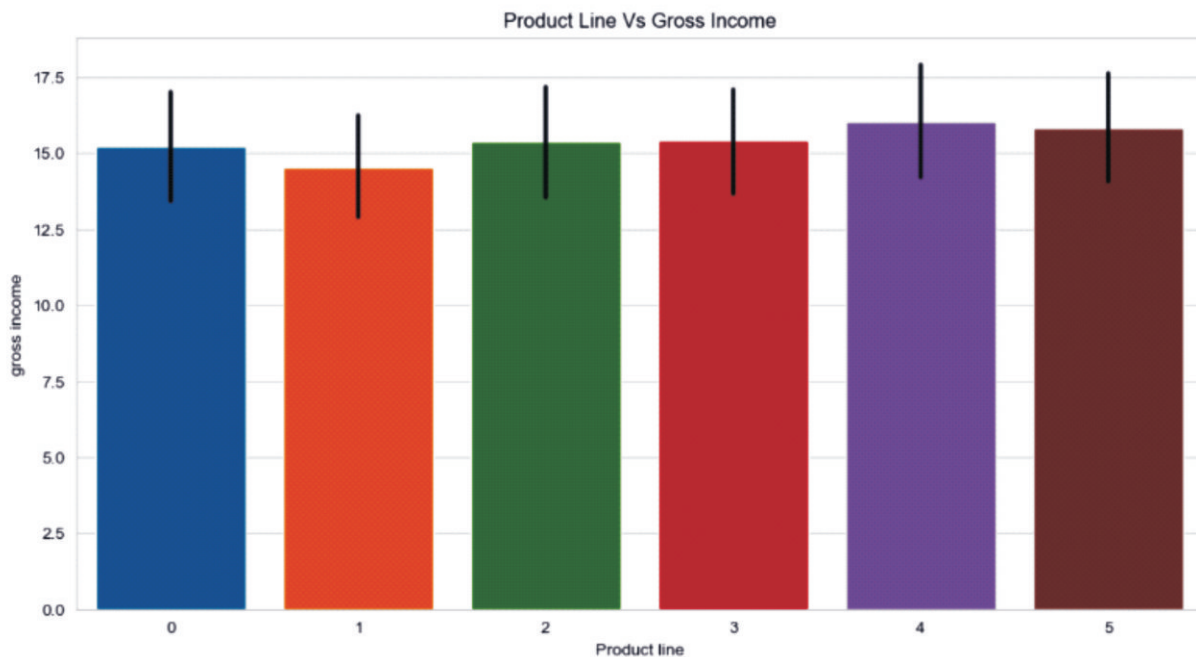


Figure 4. Bar-Plot Between Product Line and Gross Income

ML algorithms to deal with hard and tough problems in everyday life. This study looks at the complexity, capabilities, and possibilities of transforming grocery sales forecasts for a new way through this article. This paper also examines how XGBoost can change the shopping experience by enabling measurable data on supermarket performance and sustainability.

5.1.2 Random Forest

Random forest is widely used for classification and regression tasks as a machine learning algorithm (Tsoumakas, 2019). It works by creating various decision trees during the training phase and results in the mean prediction (regression) or mode of the classes (classification) of separate decision trees. This study extracts a random fragment of the training dataset for each tree to train, which makes decisions according to its own set of characteristics, reducing overfitting and strengthening generalization. During forecasting, the algorithm aggregates the predictions of each tree to produce the outcome (Chu & Zhang, 2003). This ensemble approach not only improves accuracy but also provides insights into feature importance, making Random Forest an effective tool for various predictive modeling tasks, particularly in complex datasets with noisy or correlated features (Ranjitha & Spandana, 2021). Additionally, its ability to handle large datasets with high dimensionality and maintain robustness against overfitting contributes to its popularity in practical applications across diverse domains.

5.1.3 Linear Regression

A linear regression method, which is based on a linear equation for the observed data, is used for models of relationships between dependent variables and one or more independent variables. This study's aim is to find the best fit line that reduces the variance between actual and predicted values. To forecast the value of the dependent variable based on the values of the independent variables, this best-fit line can be used. Linear regression aims to quantify and understand the relationship between variables, making it a valuable tool for prediction and inference in several areas such as

economics, business, and life sciences. The ease, clarity, and understandability of linear regression make it a fundamental technique in statistical analysis, serving as a basis for more complex models and techniques. The linear regression algorithm accepts a linear relationship between variables and may not be appropriate for other types of data.

5.1.4 LASSO Regression

LASSO (Least Absolute Shrinkage and Selection Operator) is a regression technique that incorporates regularization in order to improve the model's predictability and interpretability by avoiding overfitting (Isaac et al., 2012). In contrast to linear regression, LASSO regression introduces a sanction term that limits the absolute size of coefficients and thus decreases certain coefficients toward zero. This regularization encourages simpler models by encouraging sparsity, which means it likes to set the value of some coefficients to exactly zero for effective selecting variables. LASSO regression is especially useful where the number of features is greater than the number of observations dealing with high-dimensional data, as it can automatically select the most relevant attributes (features) while minimizing the impact of irrelevant ones. This study makes LASSO-Regression a powerful tool in fields like genetics, economics, and machine learning, where feature selection and interpretability are crucial (Odegua, 2020). However, it is important to fine-tune the regularization parameter to balance between bias and variance effectively, ensuring optimal model performance.

6. Results and Discussion

6.1 Performance Metrics

This study used the MAE (Mean Absolute Error) while analyzing and estimating the model. This means the model with a lower MAE is better. The performance metrics as an option that a task, like an MAE, consists of regression tasks. A good measure of model performance is the tested and reliable metrics.

6.1.1 RMSE

The most widely used statistic to evaluate accuracy is the Root Mean Square Error (RMSE). Predictive models, such

as the regression models, estimate the correctness of the model's forecasting in relation to the observed values. RMSE is used to analyze data by quantifying the average size of the error between projected and actual values. A reduction in RMSE indicates an improvement in the model's accuracy.

Root Mean Square Error can be expressed as:

$$RMSE = \sqrt{\frac{\sum_{i=1}^N \|y(i) - \hat{y}(i)\|^2}{N}}$$

Where, N is the number of data points,

y(i) is the i-th measurement,

and $\hat{y}(i)$ is its corresponding prediction.

6.1.2 R-Squared

In statistics, the coefficient of determination, denoted R^2 or r^2 and pronounced "R squared," is the proportion of the variation in the dependent variable that is predictable from the independent variable. The correlation between independent and dependent variables describes the strength of their relationship. The scale to which variance is explained by variation between two variables has been described by R^2 (R-squared). This means that if the R^2 of the model is 0.50, approximately half of the noted changes can be attributed to inputs from these models.

The formula for calculating R-squared is:

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y}_i)^2}$$

Here,

The value of R^2 can be in between (0 and 1).

$R^2: 1$, It talks about the variations in the dataset.

$R^2: 0$, No variations in the dataset.

R-square gives information about a model to be good-fit but not necessarily provides good-fit.

Table 2 shows the comparison of error measurements and R squared. This study has observed that with an R-squared

Algorithm	RMSE	Rsquared
XG-Boost	0.187058657378	0.999753822298
Random Forest	0.741194755243	0.999961349264
LASSO	0.160394667853	0.999999877675

Table 2. Comparison of Error Measurements

of 0.608451, the XG algorithm performs best among all three.

Conclusion

In order to forecast sales of different supermarket products, this study provides information on the basic principles of machine learning and data processing and modeling methods. More can be used successfully and innovatively to anticipate sales, using many occurrence parameters and several other elements. Accuracy is essential to predicting systems, and if the parameters used are increased, this may lead to significant improvements. Moreover, the productivity of the system could be increased by the way the submodels function.

The big retailers are striving for precise forecasts in order to avoid losses because the accuracy of their sales estimates has a direct impact on profitability. This study developed a model for sales prediction of supermarket data using the XGBoost method. The sales data from the supermarket for a given outlet to estimate product sales. The experiments demonstrate that, compared to alternative methods, this approach of XGBoost results in more accurate estimates.

References

- [1]. Abhinaya, C. S. D., Lahari, B., Priya, C. D., Anjali, D., Navya, B. S., & Jyothi, B. S. (2023). Supermarket sales prediction using machine learning. *EPRA International Journal of Research and Development (IJRD)*, 8(11), 1-6. <https://doi.org/10.36713/epra2016>
- [2]. Chu, C. W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics*, 86(3), 217-231. [https://doi.org/10.1016/S0925-5273\(03\)00068-9](https://doi.org/10.1016/S0925-5273(03)00068-9)
- [3]. Cumby, C., Fano, A., Ghani, R., & Krema, M. (2004, August). Predicting customer shopping lists from point-of-sale purchase data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 402-409). <https://doi.org/10.1145/1014052.1014098>
- [4]. Isaac, O. A., Adedeji, A. A., & Ismail, I. R. (2012).

Polynomial regression model of making cost prediction in mixed cost analysis. *International Journal on Mathematical Theory and Modeling*, 2(2), 14-23.

[5]. Kuo, R. J., Hu, T. L., & Chen, Z. Y. (2009, February). Application of radial basis function neural network for sales forecasting. In *2009 International Asia Conference on Informatics in Control, Automation and Robotics* (pp. 325-328). IEEE.

<https://doi.org/10.1109/CAR.2009.97>

[6]. Niu, Y. (2020, October). Walmart sales forecasting using XGBoost algorithm and feature engineering. In *2020 International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)* (pp. 458-461). IEEE.

<https://doi.org/10.1109/ICBASE51474.2020.00103>

[7]. Nunnari, G., & Nunnari, V. (2017, July). Forecasting monthly sales retail time series: A case study. In *2017 IEEE 19th Conference on Business Informatics (CBI)* (Vol. 1, pp. 1-6). IEEE.

<https://doi.org/10.1109/CBI.2017.57>

[8]. Odegua, R. (2020). *Applied Machine Learning for Supermarket Sales Prediction*. Project: Predictive Machine Learning in Industry.

[9]. Raizada, S., & Saini, J. R. (2021). Comparative analysis of supervised machine learning techniques for sales forecasting. *International Journal of Advanced*

Computer Science and Applications, 12(11), 102-110.

[10]. Ranjitha, P., & Spandana, M. (2021, May). Predictive analysis for big mart sales using machine learning algorithms. In *2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 1416-1421). IEEE.

<https://doi.org/10.1109/ICICCS51141.2021.9432109>

[11]. Silva, A. L., & Cardoso, M. G. (2005). Predicting supermarket sales: The use of regression trees. *Journal of Targeting, Measurement and Analysis for Marketing*, 13, 239-249.

<https://doi.org/10.1057/palgrave.jt.5740150>

[12]. Suma, V., & Hills, S. M. (2020). Data mining based prediction of demand in Indian market for refurbished electronics. *Journal of Soft Computing Paradigm (JSCP)*, 2(02), 101-110.

<https://doi.org/10.36548/jscp.2020.2.007>

[13]. Tsumakas, G. (2019). A survey of machine learning techniques for food sales prediction. *Artificial Intelligence Review*, 52(1), 441-447.

<https://doi.org/10.1007/s10462-018-9637-z>

[14]. Wang, H. (2019). Sustainable development and management in consumer electronics using soft computation. *Journal of Soft Computing Paradigm (JSCP)*, 1(01), 56.

<https://doi.org/10.36548/jscp.2019.1.006>

ABOUT THE AUTHORS

Likita H. L., Shri Shankaracharya Engineering College, Bhilai, Chhattisgarh, India.

Hireesh Yadav, Shri Shankaracharya Engineering College, Bhilai, Chhattisgarh, India.

Anjali Hedao, Shri Shankaracharya Engineering College, Bhilai, Chhattisgarh, India.

Swagata Bera, Shri Shankaracharya Engineering College, Bhilai, Chhattisgarh, India.

Siddhartha Choubey, Shri Shankaracharya Engineering College, Bhilai, Chhattisgarh, India.

ROLE OF ARTIFICIAL INTELLIGENCE IN INVESTMENT MANAGEMENT

By

MAHIBA G. B. *

JESINTHA P. **

*-** Department of Commerce, Women's Christian College, Nagercoil, Tamil Nadu, India.

<https://doi.org/10.26634/jds.2.2.21268>

Date Received: 14/10/2024

Date Revised: 26/10/2024

Date Accepted: 07/11/2024

ABSTRACT

AI is transforming the field of investment management by enhancing decision-making, improving operational efficiency, and optimizing portfolio management. As the technology evolves, it is expected that AI will play an even more integral role in driving the future of investment management, offering both opportunities and challenges for traditional asset managers. This study aims to know the most preferred artificial intelligence investing apps in investing management and to study the awareness among investors in artificial intelligence investing management. For this study, 120 investors were selected from Kanniyakumari district using the convenience sampling method. Primary and secondary data were collected for this study, and SPSS tools were used to analyze the data. This study found that there is no significant difference between gender and the awareness among investors of AI in investing management. This study concludes that AI plays a transformative role in investment management by enhancing decision-making, optimizing portfolios, and enabling personalized strategies, ultimately driving more efficient and effective investment outcomes in an increasingly complex financial landscape.

Keywords: Artificial Intelligence, Investing Management, Investing Apps, Optimizing Portfolios, Effective Investment.

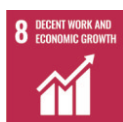
INTRODUCTION

Artificial intelligence plays a significant role in investment management, as it analyzes sets of data and can use this new set of information to generate stock market predictions. This ranges from market trends and prices of various stocks to the performance of key investments. By utilizing machine-learning algorithms, AI systems can uncover patterns and relationships that may be difficult for human investors to identify. This enables investors to make informed and data-driven decisions (Karthika, 2023). The integration of AI into investment management is not without challenges. Issues such as model transparency, ethical considerations, and regulatory compliance must be carefully managed to ensure that AI-driven decisions are fair, responsible, and aligned with

investors' goals. With the turbulent, fluctuating financial circumstances faced by the world, investors are more leaning towards long-term investments. These transitional times are presenting investors with numerous investment options like bonds, mutual funds, saving schemes, stocks, gold, etc. Many people have no idea about how to invest and where. The advanced technology, i.e., artificial intelligence, provides valuable information about various stock options and previous years investment performances, which will help the investors identify potential investment opportunities (Lalitha, 2024). AI is transforming the field of investment management by enhancing decision-making, improving operational efficiency, and optimizing portfolio management. As financial markets grow more complex, the ability to process vast amounts of data quickly and accurately has become critical. As AI continues to evolve, it is essential for investment managers to balance the benefits of these advanced technologies with a responsible approach that addresses potential risks.



This paper has objectives related to SDGs



1. Literature Review

Liu (2023) investigated the investment principles of enterprise annuities and further studied the investment management of enterprise annuities. The study results show that the strength of preferential tax policies is insufficient, and by increasing the types of portfolios, it enriches the choices of employees. It suggested that the central and local governments should perfect the legal system to provide legal guarantee for the development of enterprise annuities. The study concluded that in order to promote the development of enterprise annuities, strengthen the investment management of enterprise annuities, and improve the income of enterprise annuities.

Cao (2024) aimed to safeguard against AI exploitation in e-commerce investment management. The results show that the possible abuses of AI in e-commerce investment include market manipulation, fraud, front-running, data overfitting, and market impact. The study concluded that regarding the "White Paper on AI" and the establishment of an ecosystem of excellence to facilitate the development and implementation of artificial intelligence in the economy and public administration throughout the European Union, creating a trustworthy ecosystem should help guide the ethics of reliable AI.

2. Statement of Problem

The rapid integration of Artificial Intelligence (AI) into investment management has revolutionized the industry, offering unprecedented opportunities for enhanced decision-making, efficiency, and profitability. This technological advancement presents significant challenges. As a result, it is essential to explore how AI can be effectively leveraged to enhance investment outcomes while addressing potential challenges, these concerns and ensuring that both institutional and individual investors can benefit from its potential. This study aims to analyze the role of AI in investment management comprehensively.

3. Objectives of the Study

- To study the awareness among investors of AI in investing management.

- To know the most preferred AI investing apps in investing management.

4. Research Methodology

In consideration of the objectives, the methodology has been adopted for conducting this study. Respondents have been selected from the population as per convenience. Data has been collected from 120 respondents from Nagercoil City, Tamil Nadu, India. The sampling method used in this study was convenience sampling.

4.1 Data Source

This study is both descriptive and analytical in nature, and hence both primary and secondary data have been used for this study.

4.2 Socio-Economic Status of the Respondents

Table 1 shows the socio-economic status of the respondents, such as gender, age, educational qualification, monthly income, marital status, nature of family, and area of residence.

Table 1 shows that the majority of the respondents gender is male (70 percent). The majority of the respondents were in the age group of 41–60 years (30.8 percent). 39.2 percent of the respondents were postgraduates. It also shows that 36.7 percent of the respondents earn a monthly income of Rs 41,000-Rs 60,000, 68.3 percent of the respondents were married, and 62.5 percent were from nuclear families. Most of the respondents were from urban areas, with 76.66 percent.

4.3 Awareness among Investors on AI in Investment Management

The awareness among investors about AI in investment management is driving a shift towards more data-driven, automated strategies that promise increased efficiency, reduced bias, and enhanced portfolio performance.

Table 2 shows the result of an independent sample 't' test on awareness among investors of AI in investing management. Capacity and limitations, transparency, regulatory concerns, performance monitoring, robot-advisors, portfolio optimization, and data privacy and security are the variables of awareness among investors

on AI in investing management. Since the P value is higher than 0.05, the null hypothesis is accepted at the five percent level of significance. There is no significant difference between gender and awareness among

Variables	Demographical Factors	No. of Respondents	Percentage
Gender	Male	84	70
	Female	36	30
	Total	120	100
Age	Below 20 Years	19	15.8
	21 - 40 Years	34	28.3
	41 - 60 Years	37	30.8
	61 Years and Above	30	25.0
	Total	120	100
Education	Upto HSC	13	10.8
	Graduate	41	34.2
Qualification	Post-Graduate	47	39.2
	Professional	19	15.8
	Total	120	100
	Monthly Income	Below Rs. 20,000	18
Income	Rs. 21,000-Rs. 40,000	32	26.7
	Rs. 41,000-Rs. 60,000	44	36.7
	Rs. 61,000 and Above	26	21.6
	Total	120	100
Marital Status	Married	82	68.3
	Unmarried	38	31.6
	Total	120	100
Nature of Family	Nuclear	75	62.5
	Joint	45	37.5
	Total	120	100
Area of Residence	Rural	28	23.22
	Urban	92	76.66
	Total	120	100

Source: Primary Data

Table 1. Socio-Economic Status of the Respondents

Sl.No	Awareness	Mean		T	Sig.
		Male	Female		
1	Capacity and Limitations	3.1071	3.3056	.689	.493
2	Transparency	3.5357	3.5833	.183	.856
3	Regulatory Concerns	3.4242	3.3704	.217	.829
4	Performance Monitoring	3.5357	3.6111	.301	.765
5	Robot-Advisors	3.5833	3.5000	.372	.711
6	Portfolio Optimization	3.5714	3.6111	.154	.878
7	Data Privacy and Security	3.3611	3.3611	1.379	.172

Source: Primary Data

Significant at 5 Percent Level

Table 2. Awareness among Investors

Awareness	Strongly Agree	Agree	Neutral	Disagree	Strongly Disagree	Total Score	Mean Score	Rank
Capacity and Limitations	55(275)	47(188)	4(12)	8(16)	6(6)	497	4.14	I
Transparency	42(210)	28(112)	24(72)	24(72)	5(5)	441	3.675	III
Regulatory Concerns	10(50)	36(144)	8(24)	31(62)	35(35)	315	2.625	VII
Performance Monitoring	58(290)	40(160)	6(18)	6(12)	10(10)	490	4.083	II
Robot-Advisors	23(115)	34(136)	9(27)	13(28)	41(41)	345	2.875	VI
Portfolio Optimization	32(160)	55(220)	5(15)	17(34)	11(11)	440	3.66	IV
Data Privacy and Security	34(170)	11(44)	7(21)	47(94)	21(21)	350	2.91	V

Source: Primary Data

Table 3. Awareness among Investors

investors on AI in investing management. Table 3 shows the awareness among investors. As shown in Table 2, it is understood that the respondents are aware of "capacity and limitations." It was ranked first among the 7 factors, with a mean score value of 4.14. "Performance Monitoring" was ranked second with a mean score of 4.083. This study also reports that "transparency" was ranked third with a mean score of 3.675. "Portfolio Optimization" was ranked fourth with a mean score of 3.66. "Data Privacy and Security" was ranked fifth with a 2.91 mean score. "Robot-Advisors" was ranked sixth with a mean score of 2.875. "Regulatory Concerns" has the least rank of ninth position with the mean score of 2.625.

4.4 Most Preferred AI Investing App in Investment Management

An AI investing app in investment management leverages advanced algorithms and machine learning to analyze market trends, optimize portfolios, and provide personalized investment strategies, enhancing decision-making for both novice and seasoned investors.

Table 4 shows Betterment got the first rank with a Garret mean score of 52.42. The second rank was Wealthfront with a Garret mean score of 52. The third rank was given to Robinhood with a Garret mean score of 51.82. It was followed by Ziggma, Tickeron, and Empower with fourth, fifth, and sixth rank positions with the mea score of 51.18, 50.56, and 50. Upstox got the least ranked on the most preferred AI investing app in investing management, ranked in seventh position with a mean Garret mean score of 48.77.

5. Findings

- The majority of the respondents's gender is male, and the majority of the respondents were under the age

Sl.No	Particulars	Garrett Mean Score	Rank
1	Betterment	52.42	I
2	Robinhood	51.82	III
3	Wealthfront	52	II
4	Upstox	48.77	VII
5	Ziggy	51.18	IV
6	Tickeron	50.56	V
7	Empower	50	VI

Source: Primary Data

Table 4. Most Preferred AI Investing App

group of 41–60 years, and 39.2 percent of the respondents were postgraduates. It also shows that 36.7 percent of the respondents earn a monthly income of Rs. 41,000-Rs. 60,000, 68.3 percent of the respondents were married, and 62.5 percent were from nuclear families. Most of the respondents were from urban areas, with 76.66 percent.

- This study found that there is no significant difference between gender and the awareness among investors of AI in investing management.
- Investors were highly aware about the capacity and limitations of AI in investment management, and regulatory concerns have the least awareness among investors.
- Betterment got the first rank with a Garret mean score of 52.42. The second rank was Wealthfront with a Garret mean score of 52. The third rank was given to Robinhood with a Garret mean score of 51.82. Upstox got the least ranked on the most preferred AI investing app in investing management, ranking in seventh position with a mean Garret mean score of 48.77.

6. Suggestions

- There is a chance of unlicensed investing apps, which can lead to fraud and scams, a lack of accountability, data privacy concerns, and legal risks. Therefore, it is suggested to make sure that is a legitimate site before making any investment management.
- The government must develop comprehensive AI investment training programs and promote public awareness campaigns for the public to know about the use of AI investing apps for investment management.

- AI investing apps must provide proper educational materials, transparency reporting, and collaboration with legal and regulatory experts so that it may be easy for the investors to understand and to be more aware about the regulatory concern.
- This study shows that Upstox has secured the least rank, as the investors feel technical glitches and customer support response time to be slow. Another major drawback is that, compared to other platforms, Upstox might offer fewer educational resources or tutorials. It is suggested that they must improve platform stability and enhance customer support.

Conclusion

This study explores the role of artificial intelligence in investment management. It emphasizes the importance of verifying the legitimacy of platforms before engaging in investment activities. To benefit investors, the government should develop comprehensive AI investment training programs and promote public awareness campaigns. This study also covers investor awareness and identifies the most preferred AI-driven investment apps in the field of investment management. Future research could focus on enhancing prediction techniques, supporting risk and compliance management, integrating ESG factors, democratizing access, addressing ethical concerns, enabling real-time decisions, and improving collaboration with human expertise. This study concludes that AI plays a transformative role in investment management by enhancing decision-making, optimizing portfolios, and enabling personalized strategies, ultimately driving more efficient and effective investment outcomes in a complex financial landscape.

References

- [1]. Cao, Z. (2024). Towards promoting the efficiency of enterprise investment management with Artificial Intelligence. *Highlights in Business, Economics and Management*, 31, 57-64.
<https://doi.org/10.54097/3z1dxq85>
- [2]. Lalitha, K. P. (2024). A study on the role of Artificial Intelligence in investment management. *International*

Journal of Creative Research Thoughts (JCRT), 12(1), 189-200.

[3]. Liu, R. (2023). Investment management of enterprise annuity. *Journal of Innovation and Development*, 5(3), 40-45.

<https://doi.org/10.54097/curn6wit>

[4]. R. Karthika, A. A. (2023). Artificial Intelligence in investment management, asset management and warehouse management. *i-manager's Journal on Data Science & Big Data Analytics*, 1 (1), 40 - 46.

<https://doi.org/10.26634/jds.1.1.19874>

ABOUT THE AUTHORS

Mahiba G.B. is a Research Scholar in the Department of Commerce at Women's Christian College, Nagercoil, Tamil Nadu, India. She earned her Master's in Commerce from Manonmaniam Sundaranar University and her Master's in Business Administration from Annamalai University, Tamil Nadu, India.



Dr. Jesintha P. is an Assistant Professor at Women's Christian College, Nagercoil, Tamil Nadu, India. She passed the UGC-NET examination in 2011 and has 15 years of teaching experience in reputed institutions. Her research interests include Human Resource Management, Marketing, and Consumer Behavior.



A GENERATIVE AI MODEL FOR FOREST FIRE PREDICTION AND DETECTION

By

NALLUSAMY M. *

KARTHICK S. **

VETRIVEL B. ***

*.*** Roever Engineering College, Perambalur, Tamil Nadu, India.

<https://doi.org/10.26634/jds.2.2.21062>

Date Received: 06/08/2024

Date Revised: 02/09/2024

Date Accepted: 23/09/2024

ABSTRACT

Forest fires pose significant threats to forest ecosystems, impacting humans, animals, and plants reliant on these environments. Traditional detection methods rely on handcrafted features like color, motion, and texture, yet achieving accuracy remains challenging. This study introduces a novel approach using a lightweight fire detection method employing Deep Convolution Neural Networks (DCNN), considering temporal aspects for enhanced accuracy. By leveraging DCNN, this study aims to improve forest fire detection capabilities, mitigating the devastating effects of wildfires on both natural habitats and communities. This method represents a promising advancement in the field, offering potential solutions to the ongoing challenge of timely and accurate forest fire detection.

Keywords: DCNN, Lightweight Fire Detection, Forest Fire, Fire Prediction, AI Model, Traditional Detection Methods, Natural Habitats.

INTRODUCTION

Wildfires have significantly damaged the natural environment and the global economy. The destruction of homes, businesses, and infrastructure, and damage to wildlife, forests, crops, soil erosion, and air quality have resulted in billions of dollars in financial losses. Effective wildfire detection and monitoring systems have been considered integral to future smart communities to maintain public safety, protect the environment, and preserve natural resources. In addition to utilizing traditional wildfire monitoring systems based on ground or tower observations, there has been a rising tendency toward using advanced technology such as satellite imagery, which has shown a great potential for more efficient fire management. Among alternative technologies, camera monitoring systems offer several

advantages, such as fast and easy deployment, operation convenience, and application-customized imaging with adjustable resolution, quality, and angle of view while minimizing human intervention. It is a reliable technology to assist fire management teams in more efficient and informed fire management. Other studies carried out fire detection only by color reference.

The characteristic of fire with a reddish-yellow color is considered to distinguish fire from other objects. In the Red, Green, Blue (RGB) values of the images were analyzed. The object with the highest R-value component was determined as a fire area candidate. Other color spaces can also be used as a reference for detecting fire; for example, the YCbCr color space is used. In the study, YCbCr was able to overcome the weakness of RGB. YCbCr can detect fire in images with significant changes in illumination, which is difficult for RGB. In the Early Warning System, the speed of the process is crucial in responding quickly to any detected abnormal events. Because of this requirement, the integrated fire detection module must run in realtime but still maintain performance. Therefore,



This paper has objectives related to SDGs



this study proposed a real-time fire detection based on a combination of novel color-motion features and machine learning to achieve real-time processing and high accuracy in performance. Deep learning techniques have achieved excellent results in the field of machine vision. Deep learning has the characteristics of strong learning ability, strong adaptability, and good portability (Shamta & Demir, 2024). It can discover the intricate patterns in massive data by using a series of processing layers. Therefore, an increasing number of researchers have tried to use deep learning technology in the field of fire or smoke detection and have developed and designed many algorithms (Jadon et al., 2020).

1. Related Works

Jiao et al. (2019) proposed forest fire spread prediction with spatially explicit models, which offers precise mapping of fire behavior, facilitating targeted resource allocation and evacuation planning. Additionally, these models enable proactive measures to mitigate fire risks by identifying vulnerable areas and optimizing suppression efforts.

Preeti et al. (2021) proposed building predictive models, enabling early detection and proactive mitigation strategies. These approaches offered scalability and adaptability, allowing for continuous improvement and refinement based on evolving environmental factors.

Ranjani et al. (2023) proposed wide-area coverage and real-time data acquisition, enabling rapid detection and response to fire outbreaks. Additionally, these methods offer multispectral analysis capabilities, allowing for the identification of subtle changes in vegetation health and temperature indicative of fire activity.

Bot and Borges (2022) proposed a forest fire prediction that offers insights into the influence of meteorological conditions on fire behavior, aiding in accurate forecasting and risk assessment. They enabled proactive planning and resource allocation, enhancing preparedness and response strategies to mitigate the impact of wildfires.

Zhang et al. (2022) suggested that enhancing forest fire prediction through crowdsourcing and citizen science

harnesses the power of community involvement for real-time data collection and early detection, augmenting traditional monitoring methods. The approach fosters greater public awareness and engagement, facilitating more effective wildfire management and response efforts.

2. System Model

The current forest fire system is in watch towers, and humans are made to observe the location throughout. If any fire occurs, it is reported (Yandouzi et al., 2023). However, accurate human observation may be limited by operator fatigue, time of day, time of year, and geographic location, and in a wireless sensor-based fire detection system, coverage of large areas in forests is impractical due to the requirement of regular distribution of sensors in close proximity, and also battery charge is a big challenge.

2.1 Disadvantages

- *Human Observation Towers:* These serve as critical vantage points but have limitations in forest fire prediction systems.
- *Satellite Scanning:* Forest fire prediction primarily relies on two satellites, limiting comprehensive coverage and analysis.
- *Earth Scanning Duration:* A complete scan of the Earth takes 2 days.
- *Counting Problem:* Issues arise in accurately counting and monitoring potential fire sources.
- *Lack of Fire Forecasting:* There is no forecasting capability for early fire detection.

2.2 Proposed System

- A Deep Convolutional Neural Network (DCNN) model is proposed.
- The proposed method uses a Region Proposal Network (RPN) to detect suspected regions of fire (SRoFs) and non-fire regions based on their spatial features.
- The Gray Level Co-occurrence Matrix (GLCM) is used to extract fire-related features.
- A Convolutional Neural Network (CNN) is used to

classify the images into two classes: 1 (Fire) and 0 (No Fire).

2.3 Advantages

- High level of accuracy.
- High classification accuracy with reduced false alarm rates.
- Fast response to fires.
- The location of fire is detected.
- It can be used for outdoor places which covers large area.

Forest fire prediction and detection involve data sources such as satellite imagery, weather sensors, and environmental monitoring stations (Kim & Muminov, 2023). Pre-processed data is fed into machine learning algorithms, including convolutional neural networks (CNNs) and decision trees, to train predictive models.

These models analyze historical data and current conditions to predict the likelihood and spread of forest

fires. Simultaneously, real-time data streams enable continuous monitoring and updating of the predictive models. Figure 1 shows the Forest Fire Prediction Architecture.

The detection module employs image processing techniques and machine learning algorithms to identify potential fire hotspots in satellite imagery. An alerting system notifies relevant authorities and stakeholders upon fire detection. A dashboard interface provides visualization of prediction results, aiding in decision-making and resource allocation. Cloud infrastructure supports the system's scalability and flexibility, ensuring seamless operation.

Feedback loops enable continuous learning and optimization of the predictive models. This architecture offers a comprehensive framework for proactive forest fire prediction and detection, minimizing the impact of wildfires on ecosystems and communities (Wahyono et al., 2022).

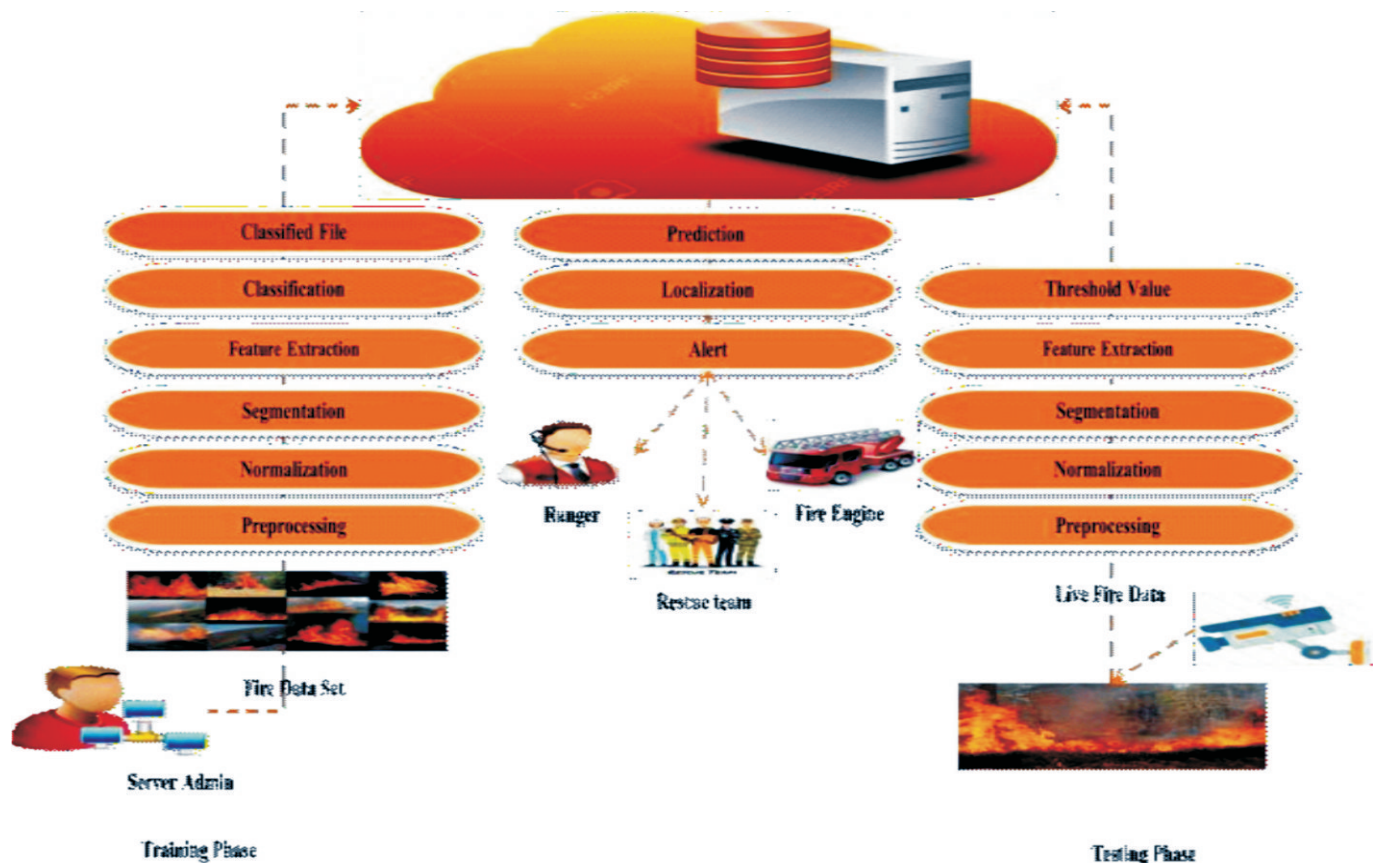


Figure 1. Forest Fire Prediction Architecture

3. Proposed Work

3.1 Fire Detect API Application

The Forest Fire Prediction System created the application portal, which provides an interface for general users, manufacturers, and the fire service, while this service serves as a link between the fire and the camera. It is the backing store for the entire system. It informs firefighters nearby of the detected fires and enables the rescue team. Figure 2 shows the fire detect API application.

3.1.1 Dataset Annotation

Dataset annotation in forest fire prediction involves labeling data points with relevant information, facilitating the training of machine learning models for accurate fire detection and prediction. Figure 3 shows the data annotation.

3.1.2 Pre-Processing

Fire image preprocessing are the steps taken to format images before they are used by model training and inference. Pre-processing steps were applied to all the raw input images to convert them into clean versions. Figure 4 shows the preprocessing.

3.1.3 Fire Detection Segmentation

Fire detection segmentation in forest fire prediction entails identifying and delineating fire-affected regions within imagery or data, enabling precise mapping and analysis for enhanced prediction and response strategies. Figure 5 shows the fire detection segmentation.

3.1.4 Feature Extraction

Feature extraction in forest fire prediction involves identifying and quantifying relevant characteristics from data sources such as satellite imagery or weather data to enhance the accuracy of predictive models. Figure 6 shows the feature extraction.

3.1.5 Classification

Classification in forest fire prediction involves categorizing data points or regions into distinct classes, such as fire and nonfried areas, using machine learning algorithms to facilitate accurate identification and prediction of fire occurrences. Figure 7 shows the classification.

3.1.6 Prediction

In this module, a live video frame will be compared with

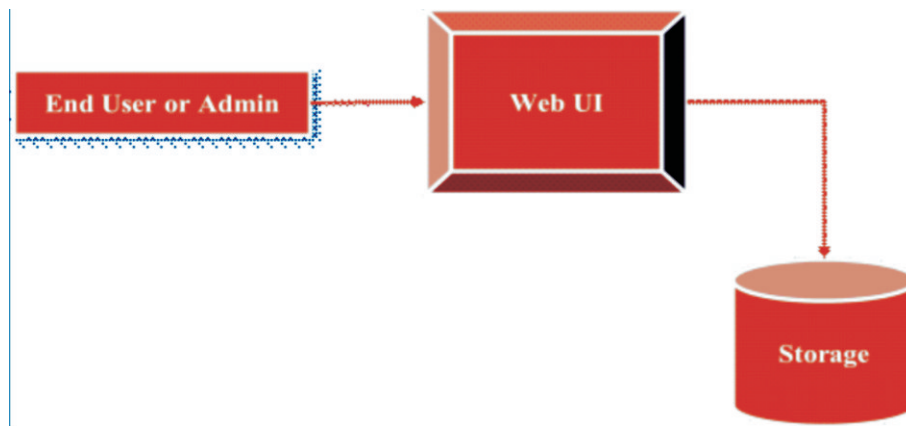


Figure 2. Fire Detect API Application

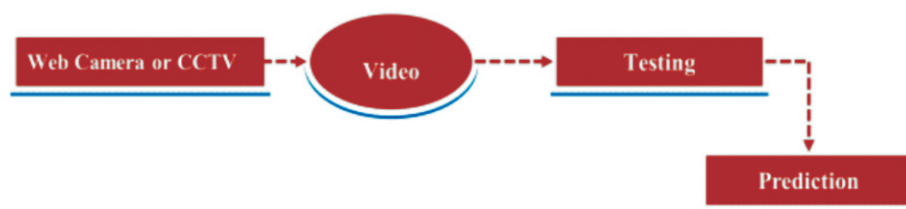


Figure 3. Data Annotation

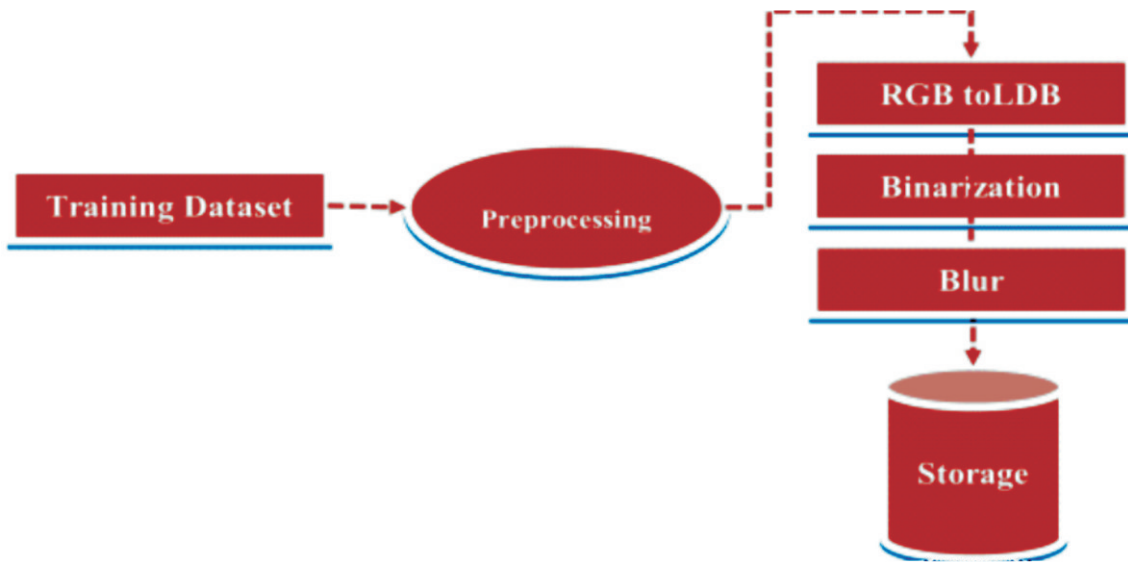


Figure 4. Pre-Processing

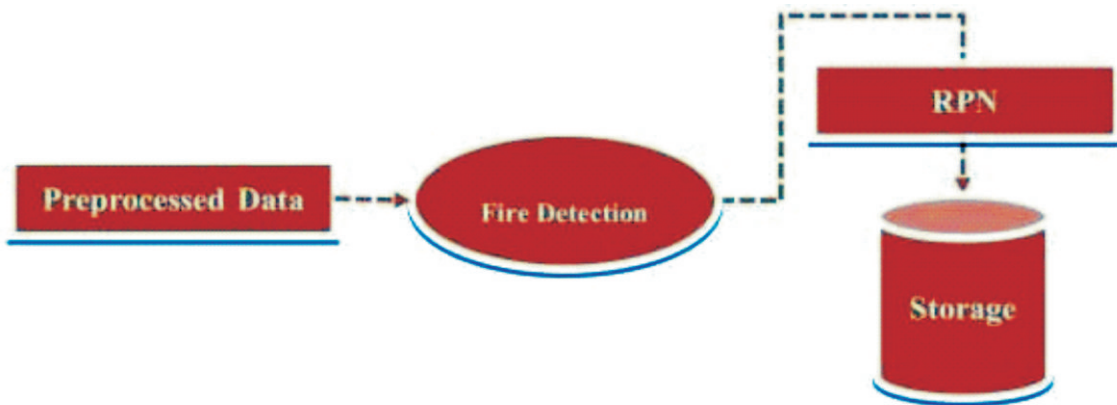


Figure 5. Fire Detection Segmentation

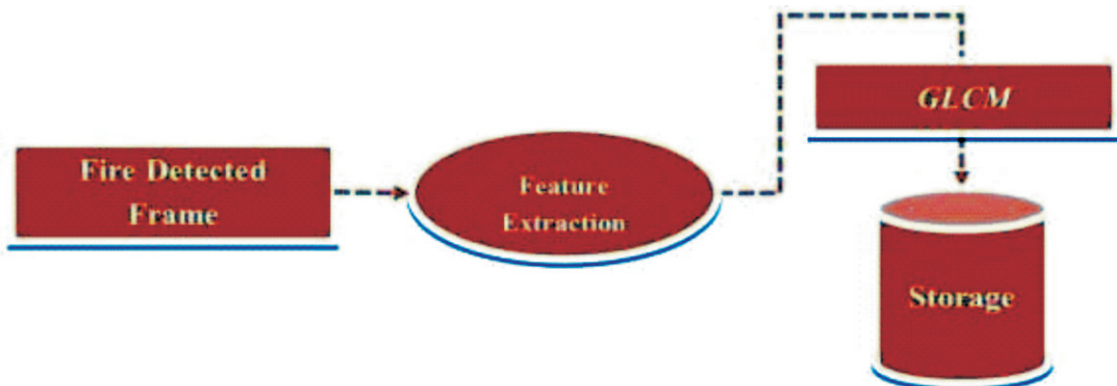


Figure 6. Feature Extraction

reference images from video of the subject for each class, and an Eigen vector is used to predict fire. Figure 8 shows the prediction.

3.1.7 Notification

Notification in forest fire prediction entails timely dissemination of alerts and warnings to relevant

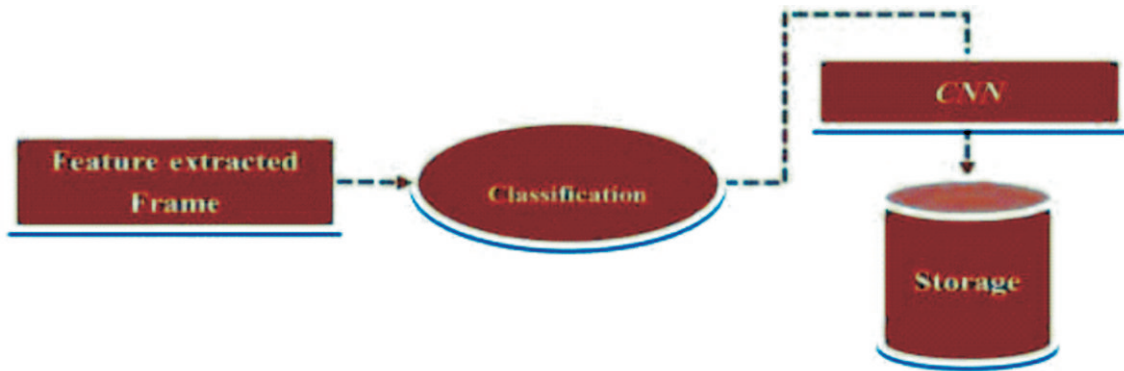


Figure 7. Classification

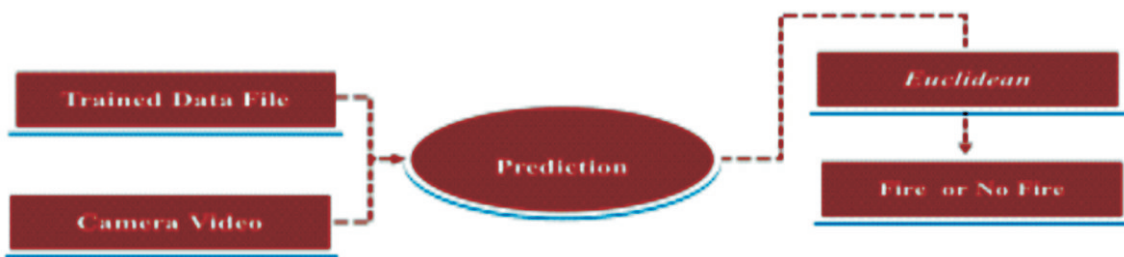


Figure 8. Prediction

authorities and communities based on predictive models and real-time monitoring, enabling prompt response and mitigation efforts. Figure 9 shows the notification.

4. Discussion

This paper introduces an innovative generative AI model for forest fire prediction and detection, primarily employing a Deep Convolutional Neural Network (DCNN) to enhance real-time monitoring and detection of wildfire risks. This proposed method uses Region Proposal Network (RPN) to detect the suspected regions of fire (SRoFs) and of non-fire based on their spatial features, and Gray-Level Co-Occurrence Matrix (GLCM) is used to extract the

features of fire. SRoFs based on spatial characteristics, increasing detection speed while maintaining high accuracy. GLCM is used for feature extraction, it enhances the model's ability to discern fire characteristics from non-fire elements, reducing false positives. This model categorizes data into fire and non-fire regions, facilitating efficient response coordination with immediate notifications to relevant authorities.

5. Future Research Directions

Feature extraction in forest fire prediction involves identifying and quantifying relevant characteristics from data sources such as satellite imagery or weather data to

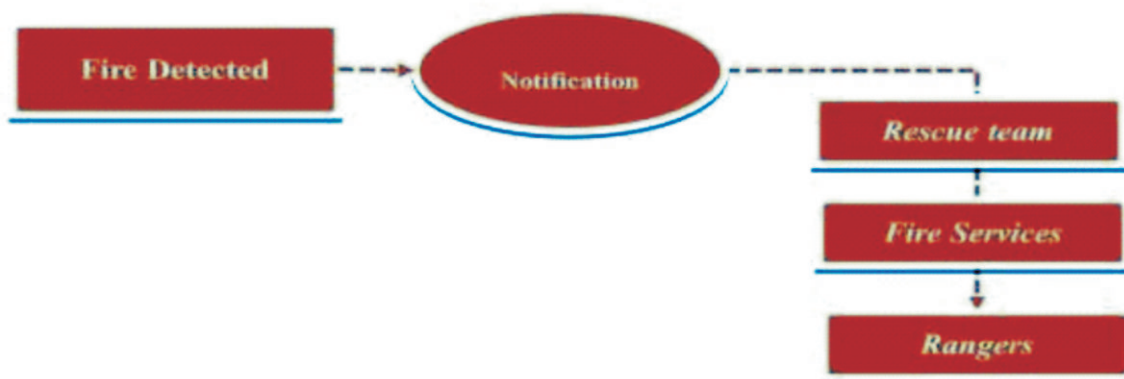


Figure 9. Notification

enhance the accuracy of predictive models.

Conclusion

There are different kinds of industries, residents, and forest areas that present specific risks to experiencing a fire incident. Addressing the fire recognition model in a multi-label approach is affordable to implement a more specific early vision-based alert system. An intelligent fire risk monitoring system is built to monitor the basic fire risk factors and obtain fire risk data in real-time. This paper presents a deep learning-based, unsupervised method for wildfire segmentation and detection. It is sensitive to small-scale fires as it reduces the omissions of forest fire point monitoring, enabling timely identification of fires in the early stage, and gaining valuable time for firefighting and rescue.

References

[1]. Bot, K., & Borges, J. G. (2022). A systematic review of applications of machine learning techniques for wildfire management decision support. *Inventions*, 7(1), 15.

<https://doi.org/10.3390/inventions7010015>

[2]. Jadon, A., Varshney, A., & Ansari, M. S. (2020). Low-complexity high-performance deep learning model for real-time low-cost embedded fire detection systems. *Procedia Computer Science*, 171, 418-426.

<https://doi.org/10.1016/j.procs.2020.04.044>

[3]. Jiao, Z., Zhang, Y., Xin, J., Mu, L., Yi, Y., Liu, H., & Liu, D. (2019, July). A deep learning based forest fire detection approach using UAV and YOLOv3. In *2019 1st International Conference on Industrial Artificial Intelligence (IAI)* (pp. 1-5). IEEE.

<https://doi.org/10.1109/ICIAI.2019.8850815>

[4]. Kim, S. Y., & Muminov, A. (2023). Forest fire smoke detection based on deep learning approaches and unmanned aerial vehicle images. *Sensors*, 23(12), 5702.

<https://doi.org/10.3390/s23125702>

[5]. Preeti, T., Kanakaraddi, S., Beelagi, A., Malagi, S., & Sudi, A. (2021, June). Forest fire prediction using machine learning techniques. In *2021 International Conference on Intelligent Technologies (CONIT)* (pp. 1-6). IEEE.

<https://doi.org/10.1109/CONIT51480.2021.9498448>

[6]. Ranjani, D., Haripriyabala, M., Indhu, J., Janani, V., & Jothi, A. (2023, April). Forest fire detection using Convolutional Neural Networks (CNN). In *2023 7th International Conference on Trends in Electronics and Informatics (ICOEI)* (pp. 852-857). IEEE.

<https://doi.org/10.1109/ICOEI56765.2023.10125611>

[7]. Shamta, I., & Demir, B. E. (2024). Development of a deep learning-based surveillance system for forest fire detection and monitoring using UAV. *Plos one*, 19(3), e0299058.

<https://doi.org/10.1371/journal.pone.0299058>

[8]. Wahyono., Harjoko, A., Dharmawan, A., Adhinata, F. D., Kosala, G., & Jo, K. H. (2022). Real-time forest fire detection framework based on artificial intelligence using color probability model and motion feature analysis. *Fire*, 5(1), 23.

<https://doi.org/10.3390/fire5010023>

[9]. Yandouzi, M., Grari, M., Berrahal, M., Idrissi, I., Moussaoui, O., Azizi, M., & Elmiad, A. K. (2023). Investigation of combining deep learning object recognition with drones for forest fire detection and monitoring. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 14(3), 377-384.

[10]. Zhang, L., Wang, M., Fu, Y., & Ding, Y. (2022). A forest fire recognition method using UAV images based on transfer learning. *Forests*, 13(7), 975.

<https://doi.org/10.3390/f13070975>

ABOUT THE AUTHORS

Nallusamy M. is working as an Assistant Professor in the Department of Computer Science and Engineering at Roever Engineering College, Perambalur, Tamil Nadu, India. He received a Master's degree in Computer Science and Engineering from Anna University, Chennai, Tamil Nadu, India. As an academician, he has published research papers in both national and international conferences. He has also published papers in journals, contributed to book chapters, and filed patents. His areas of interest include Data Analytics, Cloud Computing, Machine Learning, and Deep Learning.



Karthick S. is a Student of Computer Science and Engineering at Roever Engineering College, Perambalur, Tamil Nadu, India. During his studies, he actively participated in workshops and presented papers at symposiums and conferences. He also completed an internship in the fields of Web Design and Deep Learning.

Vetrivel B. is a Student of Computer Science and Engineering at Roever Engineering College, Perambalur, Tamil Nadu, India. During his studies, he actively participated in workshops and presented papers at symposiums and conferences. He also completed an internship in the fields of Web Design and Deep Learning.

1ST NATIONAL GRAND RESEARCH COMPETITION (NGRC 2025)

Organized by



About

i-manager Publications, in partnership with DMI Engineering College, is proud to announce the Grand Research Competition, a national-level competition designed to foster innovation and academic excellence in scientific research.

Registration link:

<https://imanagerpublications.com/NGRC2025/>

Invite for Papers

14th Oct 2024

Submission Deadline

31st Dec 2024

Judging Period

1st Jan 2025 to 15th Feb 2025

Top 100 Finalist Competition
and Award Ceremony

28th Feb 2025 on National Science Day

Finalist Competition
and Award Ceremony Venue:

DMI Engineering College, Kanyakumari Dist.



1st National Grand Research Competition (NGRC 2025)

Target Participants

- UG and PG Students
- Doctoral Candidates actively engaged in research
- Young Researchers and Academics
(From Indian institutions)

Categories of Participation

- Engineering and Technology
- Life Sciences and Health
- Environmental Sciences
- Social Sciences and Education
- Management, Business & Commerce
- Interdisciplinary Research



Awards and Prizes

First Prize: INR 1, 00,000/-

Second Prize: INR 50,000/-

Third Prize: INR 25,000/-

Top 10 Finalists: Certificates of Excellence

Top 100 would be awarded Certificate of Recognition

Publication Opportunity

Successful papers will be published in i-manager Publication's peer-reviewed scientific Journals. Participants will gain recognition in the academic community and contribute to advancing research in their respective fields.

Participation Fee: INR 250 (to be paid during submission).

Deadline: All submissions must be received by 31st DEC 2024.

+91 960600703

info@imanagerpublications.com

www.imanagerpublications.com

1ST NATIONAL GRAND RESEARCH COMPETITION (NGRC 2025)

About NGRC 2025

i-manager Publications, in partnership with DMI Engineering College, is proud to announce the Grand Research Competition, a national-level competition designed to foster innovation and academic excellence in scientific research. The competition aims to provide a platform for students, researchers, and academic professionals to showcase their research work, with successful papers being published by i-manager Publications.

Objective

The Grand Research Competition seeks to:

- Encourage new, innovative research in key scientific and technological areas.
- Provide recognition and publication opportunities for top-quality research.
- Promote collaboration and academic engagement between researchers and institutions.
- Highlight emerging research that contributes to addressing real-world problems.

Submission Guidelines

- All submissions must be in word format, with 8,000 words (including references and appendices).
- A maximum similarity tolerance of 10% will be allowed.
- Papers should be submitted through i-manager Publications' Editorial Management Platform.

Participation Fee:
INR 250
(to be paid during submission).

Deadline:
All submissions must be
received by 31st DEC 2024.

Rules & Instructions:

All entries must be submitted by 31st DEC 2024, and must adhere to the following requirements:

1. Be written and submitted by student(s) (UG, PG, Doctoral Candidates in any discipline) enrolled at an institute of Higher Education located within India.
2. The paper should have a title, a list of contributing authors with their names, contact details, designations, and affiliated institutions. Include a brief bio and a JPEG photo for each author. In case of more than one author, the corresponding authors' full details including email, fax, phone and address for communication need to be provided.
3. The paper should include an abstract (not more than 400 words), keywords, Introduction, a main body with sections

and subsections, results, discussion, conclusion, and references.

4. Be fewer than 8000 words (not including requirement #2 and not exceed 10 pages)
5. Be in English and suitable for publication (final editing assistance will be provided before publication for winning papers).
6. Be submitted in electronic format, preferably Microsoft Word or other revisable format, in Arial font, 12-point type size.
7. Be original and not previously published (copyright will remain with the author; winners grant publication rights to i-manager Publications as well as the right to release entries to other media and transfer copyright once accepted for Publication).

In addition:

- All references must be cited.
- APA format should be followed for references, citations, and for tables.
- Figures and tables should appear immediately after their first mention in the text.
- Illustrations, if provided, should be clear and in JPEG format.
- All articles should be as error-free as possible, and any technical matters must be as clear as possible.

Submission Procedure

Participants are requested to register themselves on the registration page in our Website. Once registered, authors can submit and know the status of their papers.

S.No	Judging Criteria	
1.	Originality	Does the research present new ideas or methodologies?
2.	Relevance	How significant is the research to its field and society?
3.	Research Methods	How robust are the research methodologies used?
4.	Clarity	Is the research well-organized and communicated effectively?
5.	Impact	Does the research contribute to solving real-world Challenges?
6.	UNSDG Goals	Adherence to UNSDG Goals.

i-manager's Journal on Data Science & Big Data Analytics

Published by i-manager Publications, India.

<http://www.imanagerpublications.com>

Editor-in-Chief: Dr. R. A. Thakur
*Director IITM,
India.*

Call For Papers for 2024



i-manager Publications is a leading publishing house specialized in publishing Scientific, Technology, Education and Management Journals. The credibility and impact of these publications in the Indian economic and academic environment reinforced the founding vision of i-manager Publications: promoting technology education and present international research worldwide.

i-manager's Journal on Data Science & Big Data Analytics aims to bring the best of research and innovations in the field of Data Science and Engineering and reveal the hidden patterns found in the data captured in very large databases with the precision of human intelligence. The Journal will bring bright ideas ignited in the young minds that will explore future opportunities in the area of data security, supply chain automation, industrial automation and fields that are yet to be envisioned. The Journal aims to provide a platform for academicians, researchers and students to expose the knowledge from the evolving domain of data science with fresh reviews, research experiment with new concepts in data modelling, case studies of successes in implementation of new technologies etc.

Why Publish with us?

- No Publishing Fee
- Double-blind Peer Review
- Highly qualified Editorial Board
- Publishing Journals since 2004
- Maximum Publicity in Social media
- Rapid Publication: 2-3 months

Features

Articles, Research Papers, Review Papers.

Target Audience

Academicians, practitioners and post-graduate students in the field of Data Science & Big Data Analytics, University Educational Bodies, Researchers, etc.

Submission email: submissions@imanagerpublications.com

i-manager's Journal on Data Science & Big Data Analytics

Overall Topics covered:

- Blockchain and Data Authenticity
- Stages and Components of Big Data
- Exploratory Data Analysis
- Big Data Life Cycle
- Business Intelligence
- Online Analytical Processing (OLAP) and Data Warehousing
- Cognitive Analytics
- Data Modelling and Visualization
- DevOps for Data Science
- Hadoop
- Data Mesh
- Predictive Analysis and Prescriptive Analytics
- Unstructured Data and Natural Language Processing
- System Dynamics Simulation
- Image and Video Analytics
- Econometrics
- FinTech and Modern Banking
- Probabilistic Graphical Models
- Text Mining and Analytics
- Bioinformatics
- Geospatial Technology
- Data-Centric AI
- Metadata Driven Data Fabric
- Data Visualization and Tableau



<https://www.facebook.com/imanagerPublishing/>



<https://twitter.com/imanagerpub>

Submission email: submissions@imanagerpublications.com



3/343, Hill view, Town Railway Nager, Nagercoil
Kanyakumari Dist. Pin-629 001.
Tel: +91-4652-231675, 232675, 276675

e-mail: info@imanagerpublications.com
contact@imanagerpublications.com
www.imanagerpublications.com