European Journal of
Technology
(EJT)

Framework based on Machine Learning for Lung Cancer Prognosis with Big Data-Driven

Raghuvaran Kendyala, Jagan Kurma, Jaya Vardhani Mamidala, Sunil Jacob Enokkaren, Avinash Attipalli, Varun Bitkuri

European Journal of Technology
ISSN 2520-0712 (online)
Vol.7, Issue 4, pp1 38-55, 2023

www.ajpojournals.org

Framework based on Machine Learning for Lung Cancer Prognosis with Big Data-Driven

Raghuvaran Kendyala*[1], Jagan Kurma[2], Jaya Vardhani Mamidala[3], Sunil Jacob Enokkaren[4], Avinash Attipalli[5], Varun Bitkuri[6]

[1]University of Illinois at Springfield, Department of Computer Science, [2]Christian Brothers University, Computer Information Systems, [3]University of Central Missouri, Department of Computer Science, [4]ADP, Solution Architect, [5]University of Bridgeport, Department of Computer Science, [6]Stratford University, Software Engineer,

Article history

Submitted 17.09.2023 Revised Version Received 19.10.2023 Accepted 15.11.2023

Abstract

Purpose: Lung cancer represents a fatal condition, which is typified by unregulated cell proliferation within lung tissues, which may be identified with the help of CT scans and X-rays that indicate tumors or aberrant masses. Medical imaging is essential in early diagnosis, which can enhance prognosis and the determination of an effective treatment approach.

Materials and Methods: The study presents an IQ-OTH/NCCD machine learning framework driven by Big Data to predict and prognose lung cancer based on 1,097 expert-labeled CT scan images in the dataset that runs under benign, malignant, and normal classes. The proposed system uses the preprocessing functions such as image shuffling, lung contour cropping, and resizing to utilize the inputs in the best way to train the model. To classify it, the EfficientNet-B1 is a deep learning model, a better model in accuracy and efficiency on computation thanks to the compound scaling depth, width, resolution.

Finding: Accuracy, precision, recall, and F1-score are the main performance indicators used to assess the model. The outstanding percentages of 99.10% accuracy, 99.22% precision, 97.22% recall, and 98.16% F1-scores demonstrate the model's exceptional performance. The model shows a massive improvement compared to classical models of machine learning, such as SVM or CNN. This system provides an efficient and scalable automated detection of lung cancer, thus facilitating smart healthcare with early detection and positive patient outcomes.

Unique Contribution to Theory, Practice and Policy: Future work requires a larger, multimodal dataset incorporating clinical and genetic data. Explainable AI methods should be explored to enhance generalizability. Real-world testing in smart healthcare settings and across multiple institutions is crucial for developing a practical, AI-driven tool for early lung cancer detection.

Keywords: *Lung Cancer Prediction, Machine Learning, Deep Learning, EfficientNet-B1, CT Scan Classification, Big Data, IQ-OTH/NCCD Dataset, Smart Healthcare, Image Preprocessing, Prognosis Analysis*

INTRODUCTION

In recent years, healthcare has been experiencing a paradigm shift as a result of incorporation of new and improved technologies to stimulate better diagnosis, treatment, and patient outcomes[1]. This digital transformation is reinventing the future of contemporary medicine as clinicians are able to make smart decisions and provide precision healthcare services in a more efficient manner

The emergence of big data in the medical field, which stands out for its diversity, volume, and velocity of data produced by several sources, is the main driver of this shift[2]. These are electronic health records (EHRs), diagnostic reports, sensor-based monitoring systems, laboratory findings, genomic sequences, and real-time streaming data associated with wearable technologies[3]. The compilation of these very big and heterogeneous stores offers unimaginable prospects to allow the computational models to draw meaningful conclusions to make use of clinical decision-making, disease detection, and planning of personalized treatment.

Medical imaging, particularly in the early diagnosis and prognosis, is one application of big data in the healthcare industry. X-ray, CT, MRI and PET-CT are imaging modalities that are essential in visualizing the internal anatomic structures and abnormalities[4]. In this respect, CT imaging has been found useful particularly in providing a diagnosis of thoracic disorders, such as lung disorders, since it has a high resolution with the capacity to depict minute details of lung tissues. Nevertheless, dissimilarity in the intensity changes, superimposition of anatomical structures and the errors in human interpretation causes a major problem in the correct detection of abnormalities on CT scans.

One of the most widely spread and mortal malignancies all over the world, lung cancer remains a severe threatening issue to the population because it is usually diagnosed in its late stages and also has a high rate of death. It begins in the epithelial cells in the airway and is able to spread quickly to the other organs. Although smoking is the major risk factor, non-smokers are also known to have contracted it as a result of environmental exposures and hereditary factors. Early and correct diagnosis is the key to increasing the survival rates and treatment planning. This, therefore, necessitates the increased demand for strong, scalable and data-driven lung cancer predictive and prognostic frameworks.

ML and AI have emerged as such game-changing technologies that have the potential to use healthcare big data to make precise clinical predictions. Such smart models, particularly the Deep Learning (DL) types, such as in the interpretation of medical pictures, such as pulmonary nodules and tumours found by CT scans, Convolutional Neural Networks (CNNs) have proven to be highly effective[5]. ML methods enable the discernment of obscured and non-linear patterns in data that cannot be detected by human experts, thereby facilitating more accurate diagnostic accuracy, prognosis calculation, and individualized care approaches. Incorporating big data, medical imaging, and AI technology, this paper suggests a fully-fledged ML-based model of lung cancer prediction and prognosis and provides a scalable yet smart model of early detection and treatment plan optimization.

Motivation and Contribution of Study

The program was driven by the need to improve lung cancer early detection, as it is among the primary causes of cancer-related deaths globally. The model may be trained using the IQ-OTH/NCCD dataset, which comprises expert-marked CT scan images of benign, malignant, and normal patients, in a therapeutically relevant scenario. The research develops an effective model of EfficientNet-B1 using an efficient classification method, which is thought to be accurate and efficient when used in calculations. The research will use a state-of-the-art DL

architecture coupled with a high-quality dataset to design an automated, precise, and scalable lung cancer prediction mechanism in smart healthcare systems to achieve early diagnosis and positive patient outcomes.

- Uses the IQ-OTH/NCCD dataset, ensuring clinical relevance with expert-annotated CT scans representing diverse lung cancer cases.
- A structured pre-processing approach includes image shuffling, lung contour cropping, and resizing to enhance input quality and model training effectiveness.
- The study applies EfficientNet-B1 to classify lung CT scans, compound scaling of the model's depth, breadth, and resolution to maximise performance.
- Model evaluation measures classification performance and dependability in a comprehensive manner using accuracy, recall, precision and F1-score.
- The framework supports smart healthcare by enabling efficient, automated lung cancer diagnosis for improved early detection and clinical outcomes.

Justification and Novelty of Paper

The justification for this paper lies in addressing the limitations of existing lung cancer detection methods, which often suffer from low accuracy, high computational costs, or insufficient generalization due to limited datasets or model inefficiencies. This work presents a new use of the EfficientNet-B1 model, which optimises depth, width, and resolution via compound scaling to provide better results with fewer parameters. This model makes use of the clinically validated IQ-OTH/NCCD dataset, which comprises a range of CT scans categorised as benign, malignant, and normal. The novelty further extends to the comprehensive pre-processing pipeline and balanced data handling approach, which collectively enhance model robustness and diagnostic accuracy, making the framework a valuable contribution to smart healthcare and automated lung cancer detection.

Structure of Paper

The structure of the paper is as follows In Section II, relevant work is reviewed. The dataset, pre-processing procedures, and methods are all covered in Section III. Results and analysis are presented in Section IV. Section V concludes with significant findings and suggestions for the future.

LITERATURE REVIEW

This section discusses studies using sophisticated ML approaches for predicting lung cancer. A summary of key studies is provided in Table I, highlighting various approaches applied to lung CT scan classification and prognosis within smart healthcare frameworks.

Firdaus Abdullah et al. (2020) use CT scan pictures to classify the stages of lung cancer. KNN and visual processing. Thus, the primary objective of this study is to build an image processing technique for detecting lung cancer features in CT scan images. Eliminating components from the segmented picture may facilitate the identification of lung cancer. The following procedures using image processing techniques, feature selection, data pre-processing, data collection, and lung cancer classification are all part of the desired method. A median filter was used as part of the pre-processing to eliminate any noise from the pictures. Three characteristics area, perimeter, and centroid need to be retrieved. The data set was then subjected to these attributes, which were subsequently used as inputs for the lung cancer classification scheme. The results of the investigation demonstrate the excellent accuracy of 98.15% of the KNN technique[6].

Nisha Jenipher and Radhika (2020) The methods used in ML to predict and identify LC early include input data selection, data preprocessing, feature extraction, feature selection, training, and validation, as well as selecting the best ML methodology. The second study provides a summary of the ML methods and algorithms used in LC. The accuracy, sensitivity, specificity, precision, and RMSE at different MLa are compared using the F1 score, the ROC curve, the PR curve, and the confusion matrix's AUC. The parameters utilized to create an effective ML model for the early prediction of LC are detailed in the study's conclusion[7].

Thallam et al. (2020) Any age can be affected by from young infants to the elderly, Lung cancer is one of the most prevalent and fatal illnesses worldwide. An enormous amount of money is spent each year on lung cancer diagnosis and treatment. Clinical methods like X-rays and other imaging procedures now in use are expensive and require complicated equipment. Therefore, the most important variables are the accuracy of the prediction and the use of a reliable method. This makes it necessary to develop (relatively) more economical and effective ML models for medical diagnostics utilizing medical data sets. Long-term tobacco smoking is responsible for around 85% of cases of lung cancer[8].

Sanagala et al. (2019) The experiment used the Early Lung Cancer Action Program is a publicly available dataset (ELCAP). More traditional ML methods such as SVM, k-NN, DT, RF, and others have also been used to compare the performance of the proposed CNN model. on a number of parameters, including Cohen Kappa, recall, accuracy, and precision. Additionally, the accuracy, storage capacity, and inference time of the suggested model are contrasted with those of popular CNN models like Inception V3 and VGG16. With an accuracy of 99.5%, the experimental findings demonstrate the superiority of the suggested methods over conventional ML and pre-trained models[9].

Günaydin, Günay and Şengel (2019), ML techniques to identify nodules of lung cancer, and used ANN, KNN, Naïve Bayes, Principal Component Analysis, SVM, and DT as ML techniques to identify anomalies. Every method was contrasted, both with and without preprocessing. The experimental data shows with an accuracy of 82.43% after picture processing, ANN yields the greatest results, whereas DT yields the highest results with an accuracy of 93.24% without image processing[10].

Faisal et al. (2018) aim to assess the discriminatory power of several predictors to improve the effectiveness of symptom-based lung cancer detection. ML models' classifiers are assessed using the UCI repository's benchmark dataset. It is also contrasted with popular ensembles like Majority Voting and RF. Performance assessments show that Gradient-boosted Tree performed better 90% accuracy compared to all other individual and ensemble classifiers[11].

Cengil and Çinar (2018) One of the most common illnesses worldwide is cancer. There are numerous forms of cancer. Lung cancer is the most common type of cancer. A potentially lethal illness that can affect both men and women is lung cancer. Diagnosing cancer and initiating therapy are critical to minimizing the risk of death. This article facilitates the classification of lung nodules by utilizing CT scans from SPIE-AAPM-LungX data. DL has grown in popularity as a method of categorization. It is specifically utilized for implementing the 3D convolutional neural network architecture with TensorFlow from DL frameworks[12].

Wu and Zhao (2017) Late-stage detection was the cause of mortality from lung cancer. As with other malignancies, early lung cancer identification may be necessary for the greatest chance of saving a life. This paper presents a novel neural network (NN)-based technique for small cell lung cancer (SCLC) diagnosis utilizing computed tomography (CT) images. One tactic employed is the entropy degradation method (EDM). This research might help identify lung cancers early. The National Cancer Institute's high-resolution lung CT scans were used for

training and testing. Twelve CT images of the lungs were chosen from the set; six of them showed people with SCLC, while the other six showed healthy lungs. Their model is trained using five randomly chosen scans from each group, whereas two scans are utilized for testing. Their algorithms have a 77.8% accuracy rate[13]

Table 1: Literature Summary on ML Framework for Lung Cancer Prediction

Author	Methodology	Dataset	Key Findings	Limitation/Future gap
Firdaus Abdullah et al. (2020)	Image processing and k-NN classifier; feature extraction (area, perimeter, centroid); median filtering	CT scan images	Achieved 98.15% accuracy using k-NN for lung cancer stage classification	Limited to basic features; lacks deep learning or hybrid approaches
Nisha Jenipher and Radhika (2020)	Survey and comparative analysis of ML algorithms (feature selection, classification, evaluation metrics)	Kaggle	Identified performance metrics (AUC, F1, RMSE) and effective parameters for ML models	Lacks implementation; theoretical review without empirical validation
Thallam et al. (2020)	Advocated ML models for affordable diagnosis; general discussion on tobacco's role in LC	Kaggle	Emphasized cost-effective ML-based alternatives to traditional imaging	No specific models or experimental results provided
Sanagala et al. (2019)	CNN-based model vs traditional ML (SVM, k-NN, RF, etc.); compared with VGG16 and InceptionV3	ELCAP dataset	CNN model outperformed others with 99.5% accuracy	Hardware requirements, and possible generalizability issues not addressed
Günaydin et al. (2019)	PCA + ML algorithms (ANN, SVM, k-NN, etc.); with and without preprocessing	Not specified	ANN gave best result (82.43%) after preprocessing; Decision Tree best without (93.24%)	Moderate accuracy; lacks integration with advanced DL models
Faisal et al. (2018)	Evaluated SVM, C4.5, MLP, NB, RF, Majority Voting, and GBT on symptoms	UCI repository dataset	Gradient-Boosted Trees achieved highest accuracy (90%)	Focused only on symptom-based data, not imaging or hybrid methods
CENGİL and ÇINAR (2018)	Used TensorFlow and 3D CNN on CT images	SPIE-AAPM-LungX dataset	Demonstrated effectiveness of 3D CNNs in lung nodule classification	Future work could focus on larger datasets and ensemble DL models
Wu and Zhao (2017)	An Entropy Degradation Method (EDM) for CT images based on neural networks is proposed.	NCI lung CT dataset (12 samples)	Accuracy of 77.8% in detecting SCLC	Small dataset, lower accuracy, and limited generalizability

MATERIALS AND METHODS

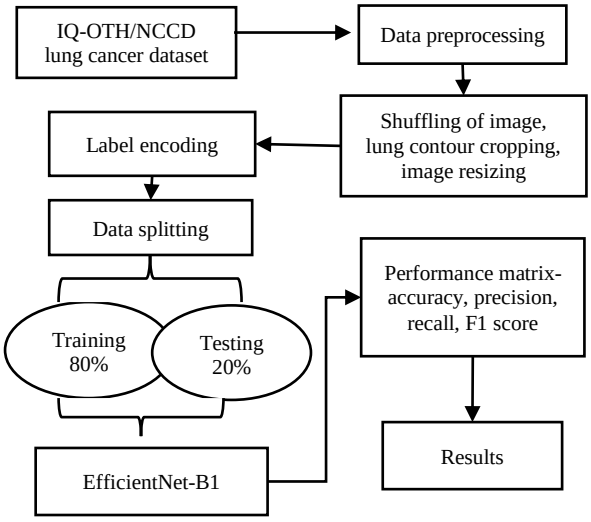

Figure 1: Flowchart for ML Framework for Lung Cancer Prediction

These flowchart steps are discussed below:

The suggested approach describes a thorough ML framework for predicting lung cancer using the EfficientNet-B1 model. The process begins with data collection from the openly accessible IQ-OTH/NCCD dataset includes 1,097 annotated CT scan pictures that have been categorized as normal, malignant, and benign. Following this, extensive preprocessing techniques are used, including enlarging the photos to 240×240×3 to fit the model, clipping the lung contour to choose the interests, and shuffling to prevent bias. The dataset is still in class-balanced shape even after class names are encoded in numerical codes and an 80:20 train. After training, the EfficientNet-B1 model may be employed with the advantages of compound scaling for performance in depth, breadth, and resolution. The confusion matrix is utilized to compute the model's typical performance measures, including F1-score, recall, accuracy, and precision. Sources of these measurements are repeatable and useful when determining the efficiency of the model and when comparing the model to other diagnostic systems that detect intelligent lung cancer in the background of smart health care. Figure 1 shows the Lung Cancer Prediction Flowchart using ML.

Data Collection

The aforementioned specialized hospitals acquired the lung cancer dataset collected over three months in the autumn of 2019 from the Iraq-Oncology Teaching Hospital/National Centre for Cancer Diseases (IQ-OTH/NCCD). Included are CT images of both healthy individuals and patients with lung cancer at various stages. The IQ-OTH/NCCD slides were annotated by radiologists and oncologists from both locations. As shown in Figure 2, the collection comprises 1190 pictures, which are slices of CT scans from 110 individuals. These conditions can be classified as normal, malignant, or benign. Forty of these instances have been identified as malignant, fifteen as benign, and forty as normal. The CT scans were collected using the original DICOM format, and each picture had many slices. These slices, which vary in quantity from 80 to 200, each present the human chest from a distinct perspective.

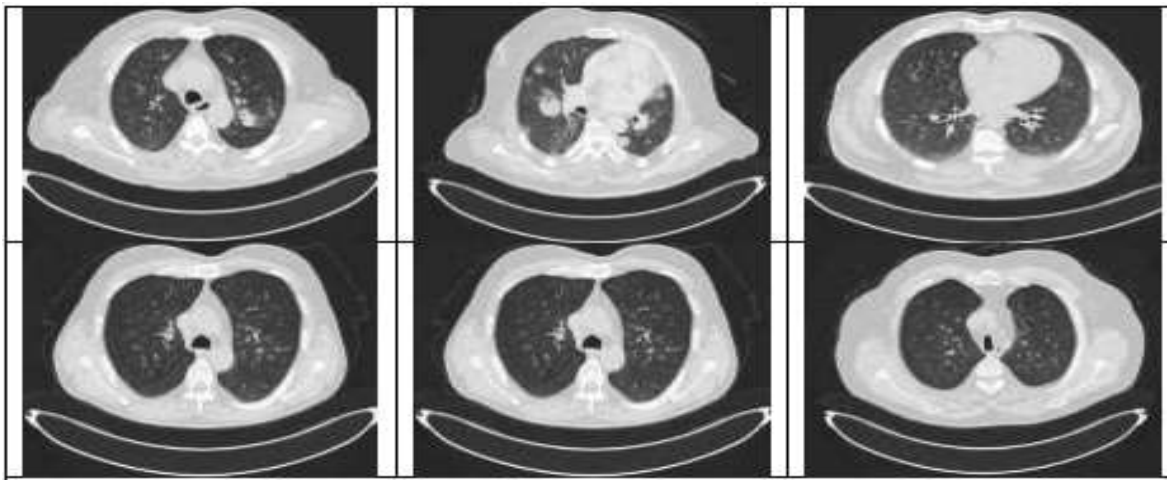

Figure 2: CT Scan Images from the IQ-OTH/NCCD Dataset

Data Pre-processing:

Standard ML models and intrusion detection boosting classifiers require pre-processing of raw data. To make sure the input data was ready for the training models, this study used a variety of pre-processing strategies and assessments, prior to completing the main pre-processing processes for data cleaning, which include eliminating superfluous elements like the traffic flow recording time and date, managing missing and infinite values.

- **Shuffling of Images:** All images from each class (benign, malignant, and normal) are shuffled to avoid any bias during training.
- **Lung Contour Cropping:** To reduce noise and focus on relevant regions, each CT scan was cropped around the largest lung contour, removing background and retaining only the essential lung area for analysis.
- **Image Resizing:** After cropping, all images were resized to 240 × 240 × 3 to match the input requirements of the EfficientNetB1 model. This resizing helps maintain important contextual information while reducing computational complexity during training.

Label Encoding

Class labels were numerically encoded as 0 for benign, 1 for malignant, and 2 for normal to make them compatible with the classification model, enabling it to effectively interpret and learn from the target classes.

Data Splitting

To create an objective model training and assessment distribution, each class is randomised at random, and the dataset is divided 80:20 between training and testing sets.

Proposed Efficientnet-B1 Model

The EfficientNet-B1 Model is predicated on straightforward and incredibly successful compound scaling techniques. For transfer learning datasets, this technique allows a baseline ConvNet to be scaled up to any goal resource limits while preserving model efficiency.[14]. In general, the Efficient Net model performs more accurately and efficiently than the present CNNs. Better results are obtained by Efficient Net by scaling down the model equally in depth, breadth, and resolution. Between B1 and the other seven models, the number of parameters remains constant as the accuracy of the model increases. Efficient Net B1, the model that forms the basis of all subsequent Efficient Net models, is schematically depicted in Figure 3.

European Journal of Technology
ISSN 2520-0712 (online)
Vol.7, Issue 4, pp1 38-55, 2023

www.ajpojournals.org

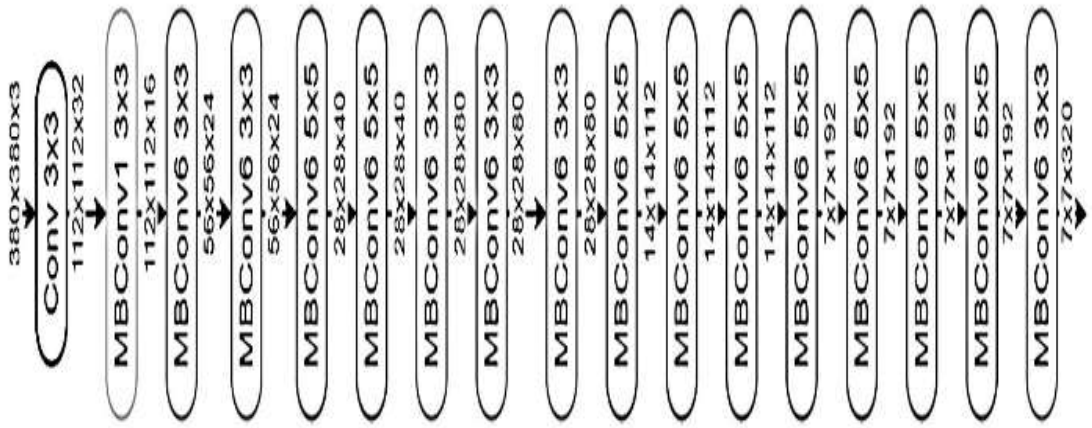

Figure 3: EfficientNet-B1 Architecture

A CNN has three scaling dimensions: depth, breadth, and resolution. A network's depth, which is equal to its width, is determined by the number of layers that comprise it. This only serves as an illustration of the CNN network. The word "resolution" refers to the image quality that is transmitted to CNN. To scale network depth, breadth, and resolution equally,Efficient Net employs a straightforward and efficient scaling method that makes use of a compound coefficient, as shown in Equations (1), (2), and (3), respectively.

$$depth \rightarrow d = \alpha^{\varphi},\ \alpha \geq 1, \quad (1)$$

$$width \rightarrow w = \beta^{\varphi},\ \beta \geq 1; \quad (2)$$

$$resolution \rightarrow w = \gamma^{\varphi},\ \gamma \geq 1; \quad (3)$$

Where the previously indicated, method is used to calculate the constants α, β, and γ.

While α, β, and γ govern how these resources are allocated to the resolution, breadth, and depth of the network, respectively, φ indicates a user-defined coefficient that indicates the quantity of resources available. The baseline is scaled using the compound scaling technique. Two-stage EfficientNet-B1 system:

- Assuming twice as many resources are available, let φ = 1. Then, do a grid search for α, β, and γ.
- Assign constants α, β, and γ based on the values established in the preceding step, then experiment with various φ values. Variations in φφ result in an Efficient Nets B1Performance matrix.

Performance Metrics

A variety of metrics are used to evaluate the efficacy of the DL algorithms created for the detection and classification of lung nodules[15]. Receiver Operating Characteristic (ROC), F1-score, recall, accuracy, and precision are displayed in the graphs. Below is a discussion of the various measures:

- **True Positive (TP):** Accurately anticipated favorable examples.
- **True Negative (TN):** Accurately anticipated negative cases.
- **False Positive (FP):** Inaccurately predicting negative cases as positive.
- **False Negative (FN):** Positive cases that were mistakenly forecast as negative.

Accuracy

The number of TP and TN findings divided by the total number of outcomes is how one determines the precision of a classifier or diagnostic test, Equation (4) can be used to express it.

$$Accuracy = \frac{TN + TP}{TP + TN + FP + FN} \quad \text{(4)}$$

Precision

Precision, which is determined by Equation (5), is the percentage of accurate positive outcomes.

$$Precision = \frac{TP}{TP+FP} \quad \text{(5)}$$

i. Recall

In Equation (6), recall is the percentage of properly identified TP instances.

$$Recall = \frac{TP}{TP+FN} \quad \text{(6)}$$

ii. F1 Score

The F1 score is determined by applying Equation (7) to the balanced harmonic mean of accuracy and recall.

$$F1 = \frac{2*(precision*recall)}{precision+recall} \quad \text{(7)}$$

iii. Loss

A statistic known as loss is used to quantify the discrepancy between the actual class labels and the expected probability; in binary classification issues, this is especially true for binary cross-entropy loss. Equation (8) is used to express it.

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \quad \text{(8)}$$

Where N is the number of samples, yi is the actual label, and $\hat{y}_i$ is the anticipated probability.

iv. Receiver Operating Characteristic (ROC) Curve

The performance of a classifier is also commonly evaluated using a ROC curve diagram. In particular, the figure is made by comparing the FPR at different threshold levels with the genuine positive rate (recall). These performance matrices are included in the IIDS for ML Framework for Lung Cancer Prediction model performance evaluation and comparison analysis.

FINDINGS

The results of the proposed AI system for predicting lung cancer based on EfficientNet-B1 are shown in this section, along with a comparative analysis against baseline models. The research employed the IQ-OTH/NCCD lung cancer dataset. The system was implemented on a high-performance computing setup comprising an Intel Core i3 processor, NVIDIA RTX 4090 GPU with 32GB VRAM, operating on Windows 10. Table II summarizes the EfficientNet-B1 model's performance on the IQ-OTH/NCCD lung cancer dataset using four crucial evaluation metrics. The model's great overall categorization capabilities are demonstrated by its 99.10% accuracy rate. It attains a precision of 99.22%, demonstrating its effectiveness in minimizing FP predictions. The model has a recall of 97.22% and is therefore effective at identifying the real cases of lung cancer, thus diminishing the chances of FN. In addition, there is a well-balanced performance between recall and accuracy, as seen by the 98.16% F1-score.

Table 2: Findings of Efficientnet-B1 Models on Iq-Oth/Nccd Lung Cancer Dataset

Performance Metric	EfficientNet-B1
Accuracy	99.10
Precision	99.22
Recall	97.22
F1-score	98.16

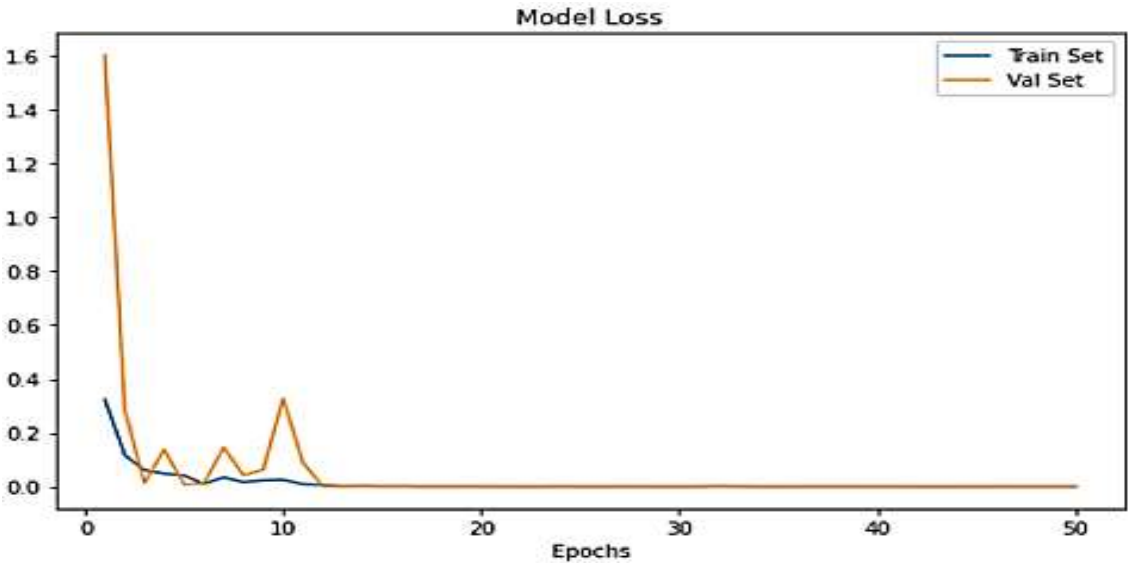

Figure 4: Loss curve of EfficientNet-B1

Figure 4 displays the loss curve following the addition of data from the sets used for training and validating the model throughout the process. As seen by the notable reduction in loss during the initial epochs, the model quickly picks up important patterns from the enriched data. The validation loss (orange line) swings initially but also converges towards 0 after around 15 epochs, in contrast to the training loss (blue line), which climbs steadily and stabilises at 0. The model's capacity to generalise has been enhanced via data augmentation, which may assist in avoiding overfitting and increase performance on unknown data.

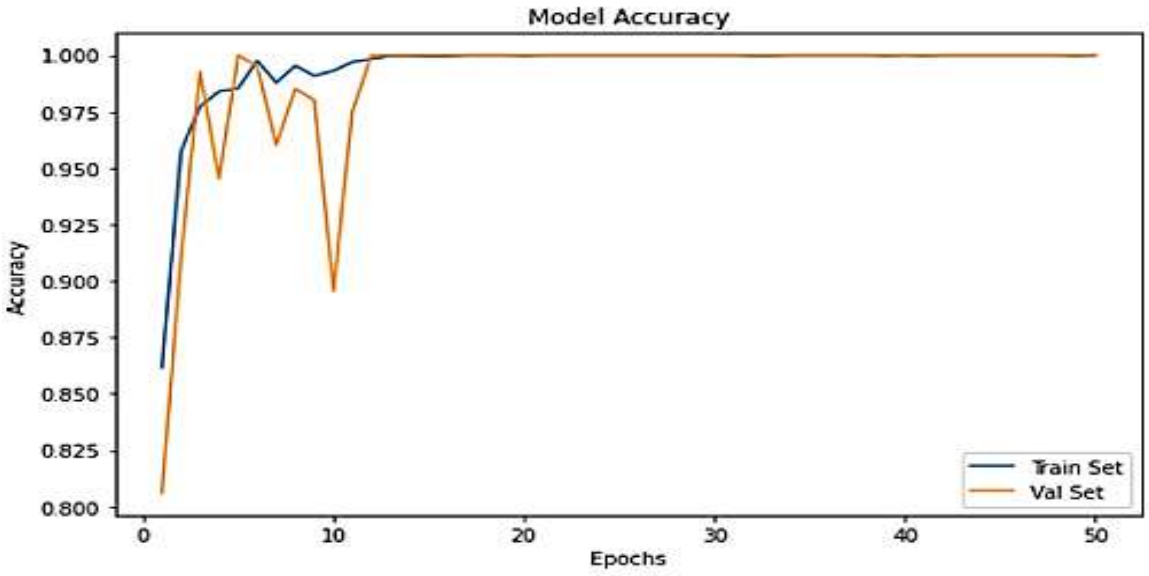

Figure 5: Accuracy Curve of EfficientNet-B1.

Figure 5 displays the accuracy curve after data augmentation for both the training and validation sets. The blue line represents the training accuracy, which increases dramatically and then levels off at almost 100% throughout the first ten epochs, demonstrating how quickly the model picks up knowledge from the richer input. The validation accuracy (orange line) initially shows some fluctuation, which is typical due to the variability introduced by augmentation, but it eventually converges and remains consistently high beyond 15 epochs.

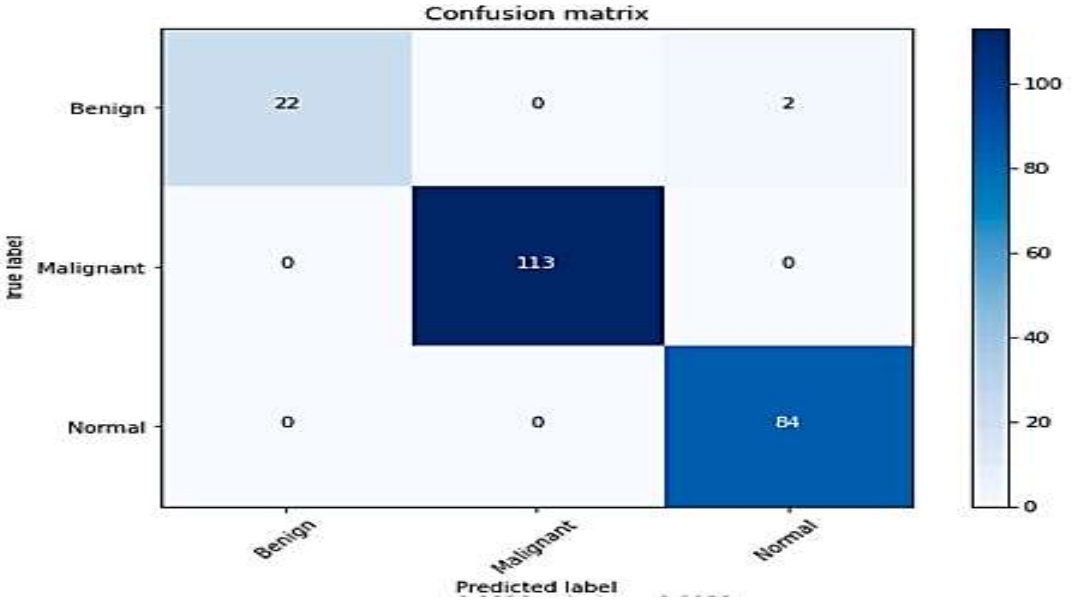

Figure 6: Confusion Matrix of Efficientnet-B1

The confusion matrix produced by the classification model trained with data augmentation is shown in Figure 6, demonstrating outstanding overall performance. The matrix indicates near-perfect classification of all three classes: Benign, Malignant, and Normal. Out of 24 benign samples, 22 were correctly identified, with only 2 misclassified as normal. All 113 malignant cases were correctly classified with zero errors, highlighting the model's strength in detecting critical cases. Similarly, all 84 normal samples were accurately classified.

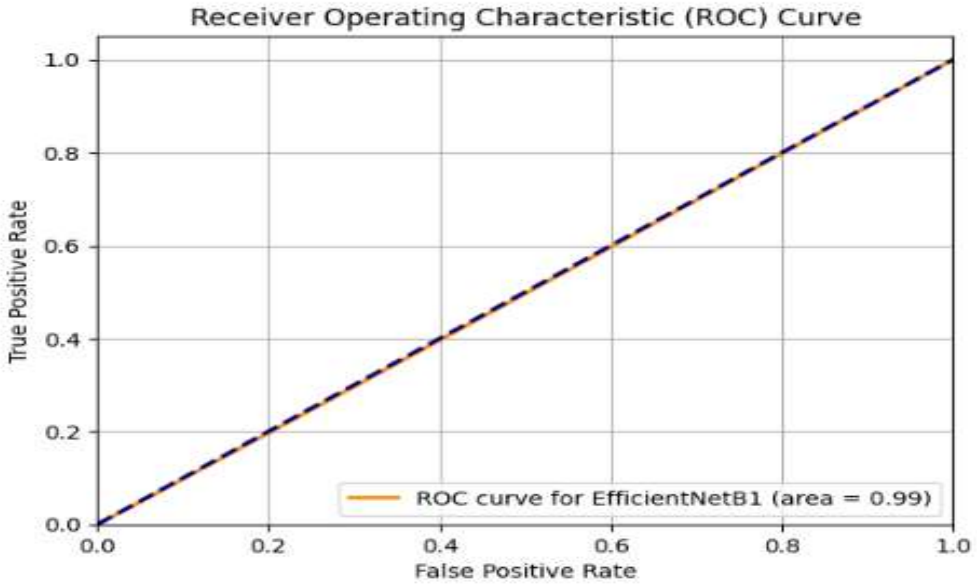

Figure 7: ROC Curve of EfficientNet-B1

The EfficientNet-B1 model's classification performance is displayed using its in Figure 7, It displays the Receiver Operating Characteristic (ROC) curve. The graph compares the FPR (1-Specificity) against the TPR (Sensitivity) at various threshold settings. As a baseline, a random classifier is represented by a dashed diagonal line. With its ROC curve closely following the upper-left corner of the Figure, the EfficientNet-B1 model exhibits outstanding performance, showing a low FPR and a high TPR at every threshold. The EfficientNet-B1 model's Area Under the Curve (AUC) of 0.99 indicates that it can detect positive and negative classes almost perfectly.

Comparative Analysis

A comparison between the proposed EfficientNet-B1-based model and the existing SVM and CNN models for lung cancer prediction is shown in Table III. The EfficientNet-B1 model significantly outperforms both benchmarks across all performance metrics. It outperforms CNN (93.54%) and SVM (80%) with the highest possible accuracy of 99.10%. In terms of memory and accuracy, EfficientNet-B1 records 99.22% and 97.22%, respectively, indicating both high correctness and sensitivity, while the CNN shows slightly lower values and the SVM lags notably. EfficientNet-B1 is also supported by the balanced accuracy and recall F1-score, which stands at 98.16%, whereas CNN and SVM achieve 95% and 69%, respectively. This result demonstrates the power and superior predictive abilities of the suggested model in detecting lung cancer.

Table 3: Comparison between proposed model performances for ML Framework for Lung Cancer Prediction

Performance Metric	**EfficientNet-B1**	**SVM**[16]	**CNN**[17]
Accuracy	99.10	80	93.54
Precision	99.22	90	97.10
Recall	97.22	57	95.71
F1-score	98.16	69	95

The proposed EfficientNet-B1 model demonstrates outstanding performance in lung cancer prediction, achieving an accuracy of 99.10%. These outcomes support very high potential of the model to give correct identification of the positive or negative cases which lowers the number of FP and FN to a large extent. This high performance is due to the compound scaling nature of EfficientNet-B1 that uniformly scales depth, width, and resolution It makes it possible for the network to more effectively discover intricate patterns in medical imaging data. Moreover, it has a lightweight model, which increases speed of inference and decreases computation expense, making it most appropriate when implemented in clinical settings in real time.

CONCLUSION AND RECOMMENDATION

Conclusion

Lung cancer is also associated with various silent progressions, and in most cases, its initial stages are characterized by mild or no symptoms at all, which is why the imaging methods are of critical importance when early stages are to be identified. New technologies such as DL models are currently being used to support lung CT scan analysis and facilitate the proper division of types of cancer. To sum up, the suggested model based on EfficientNet-B1 shows an outstanding accuracy and robustness in lung cancer prediction. Because of its excellent precision and recall, it can distinguish between normal, malignant, and benign lung CT images. The system performed better than traditional ML models due to the combination of cutting-edge preprocessing techniques and EfficientNet-B1's compound scaling, which illustrates the prospects of the system as a stable instrument of intelligent lung cancer diagnosing.

Recommendations

In future work, a larger and more comprehensive sample can be used to expand the dataset's size and include multimodal data, e.g., clinical records and genetic data, as well as to consider explainable AI methods that would increase the generalizability of the model and its potential clinical applicability. Furthermore, testing the framework in real-time situations in smart healthcare settings and testing it on other institutions will aid in channeling an effective artificial intelligence approach for a practical diagnostic tool for lung cancer early detection.

REFERENCES

[1] S. Singamsetty, "Retinal Twins: Leveraging Binocular Symmetry with Siamese Networks for Enhanced Diabetic Retinopathy Detection," *Turkish Online J. Qual. Inq.*, vol. 11, no. 4, pp. 2843–2850, 2020, doi: 10.53555/tojqi.v11i4.10607.

[2] M Supriya and A. Deepa, "Machine learning approach on healthcare big data: a review," *Big Data Inf. Anal.*, vol. 5, no. 1, pp. 58–75, 2020, doi: 10.3934/bdia.2020005.

[3] D. M. Kasthuri and M. R. Jency, "Lung Cancer Prediction Using Machine Learning Algorithms on Big Data: Survey," *Int. J. Comput. Sci. Mob. Comput.*, vol. 9, no. 10, pp. 73–77, Oct. 2020, doi: 10.47760/IJCSMC.2020.v09i10.009.

[4] S. Makaju, P. W. C. Prasad, A. Alsadoon, A. K. Singh, and A. Elchouemi, "Lung Cancer Detection using CT Scan Images," *Procedia Comput. Sci.*, vol. 125, pp. 107–114, 2018, doi: 10.1016/j.procs.2017.12.016.

[5] V. Kolluri, "Machine Learning in Managing Healthcare Supply Chains: How Machine Learning Optimizes Supply Chains, Ensuring the Timely Availability of Medical Supplies," *J. Emerg. Technol. Innov. Res.*, vol. 3, no. 6, 2016.

[6] M. F. Abdullah, S. N. Sulaiman, M. K. Osman, N. K. A. Karim, I. L. Shuaib, and M. D. I. Alhamdu, "Classification of Lung Cancer Stages from CT Scan Images Using Image Processing and k-Nearest Neighbours," in *2020 11th IEEE Control and System Graduate Research Colloquium (ICSGRC)*, IEEE, Aug. 2020, pp. 68–72. doi: 10.1109/ICSGRC49013.2020.9232492.

[7] V. N. Jenipher and S. Radhika, "A study on early prediction of lung cancer using machine learning techniques," in *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, 2020. doi: 10.1109/ICISS49785.2020.9316064.

[8] C. Thallam, A. Peruboyina, S. S. T. Raju, and N. Sampath, "Early Stage Lung Cancer Prediction Using Various Machine Learning Techniques," in *Proceedings of the 4th International Conference on Electronics, Communication and Aerospace Technology, ICECA 2020*, 2020. doi: 10.1109/ICECA49313.2020.9297576.

[9] S. S. Sanagala, S. K. Gupta, V. K. Koppula, and M. Agarwal, "A fast and light-weight deep convolution neural network model for cancer disease identification in human lungs (s)," in *Proceedings - 18th IEEE International Conference on Machine Learning and Applications, ICMLA 2019*, 2019. doi: 10.1109/ICMLA.2019.00225.

[10] Ö. Günaydin, M. Günay, and Ö. Şengel, "Comparison of Lung Cancer Detection Algorithms," in *2019 Scientific Meeting on Electrical-Electronics & Biomedical Engineering and Computer Science (EBBT)*, 2019, pp. 1–4. doi: 10.1109/EBBT.2019.8741826.

[11] M. I. Faisal, S. Bashir, Z. S. Khan, and F. Hassan Khan, "An Evaluation of Machine Learning Classifiers and Ensembles for Early Stage Prediction of Lung Cancer," in *2018 3rd International Conference on Emerging Trends in Engineering, Sciences and Technology, ICEEST 2018*, 2018. doi: 10.1109/ICEEST.2018.8643311.

[12] E. Cengil and A. Çinar, "A Deep Learning Based Approach to Lung Cancer Identification," in *International Conference on Artificial Intelligence and Data Processing (IDAP)*, 2018, pp. 1–5. doi: 10.1109/IDAP.2018.8620723.

[13] Q. Wu and W. Zhao, "Small-Cell Lung Cancer Detection Using a Supervised Machine Learning Algorithm," in *Proceedings - 2017 International Symposium on Computer Science and Intelligent Controls, ISCSIC 2017*, 2017. doi: 10.1109/ISCSIC.2017.22.

[14] G. Marques, D. Agarwal, and I. de la Torre Díez, "Automated medical diagnosis of COVID-19 through EfficientNet convolutional neural network," *Appl. Soft Comput.*, vol. 96, Nov. 2020, doi: 10.1016/j.asoc.2020.106691.

[15] D. Riquelme and M. A. Akhloufi, "Deep Learning for Lung Cancer Nodules Detection and Classification in CT Scans," 2020. doi: 10.3390/ai1010003.

[16] N. Banerjee and S. Das, "Prediction Lung Cancer- in Machine Learning Perspective," in *2020 International Conference on Computer Science, Engineering and Applications, ICCSEA 2020*, 2020. doi: 10.1109/ICCSEA49143.2020.9132913.

[17] H. F. Al-Yasriy, M. S. Al-Husieny, F. Y. Mohsen, E. A. Khalil, and Z. S. Hassan, "Diagnosis of Lung Cancer Based on CT Scans Using CNN," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 928, no. 2, 2020, doi: 10.1088/1757-899X/928/2/022035.

[34]Sandeep Kumar, C., Srikanth Reddy, V., Ram Mohan, P., Bhavana, K., & Ajay Babu, K. (2022). Efficient Machine Learning Approaches for Intrusion Identification of DDoS Attacks in Cloud Networks. *J Contemp Edu Theo Artific Intel: JCETAI/101*.

Ajay, S., Satya Sai Krishna Mohan G, Rao, S. S., Shaunak, S. B., Krutthika, H. K., Ananda, Y. R., & Jose, J. (2018). Source Hotspot Management in a Mesh Network on Chip. In *VDAT* (pp. 619-630).

Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2020). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., & Penmetsa, M. (2022). Big Data-Driven Time Series Forecasting for Financial Market Prediction: Deep Learning Models. *Journal of Artificial Intelligence and Big Data*, *2*(1), 153–164.DOI: 10.31586/jaibd.2022.1341

Dinesh, K. (2022). Navigating the link between internet user attitudes and cybersecurity awareness in the era of phishing challenges. *International Advanced Research Journal in Science, Engineering and Technology*.

Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., & Tyagadurgam, M. S. V. Efficient Framework for Forecasting Auto Insurance Claims Utilizing Machine Learning Based Data-Driven Methodologies. *International Research Journal of Economics and Management Studies IRJEMS*, *1*(2).

Gangineni, V. N., Tyagadurgam, M. S. V., Chalasani, R., Bhumireddy, J. R., & Penmetsa, M. (2021). Strengthening Cybersecurity Governance: The Impact of Firewalls on Risk Management. *International Journal of AI, BigData, Computational and Management Studies*, *2*, 10-63282.

Gopalakrishnan Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv e-prints*, arXiv-1001.

HK, K. (2020). Design of Efficient FSM Based 3D Network on Chip Architecture. *INTERNATIONAL JOURNAL OF ENGINEERING*, *68*(10), 67-73.

Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., & Kamarthapu, B. (2021). Big Data and Predictive Analytics for Customer Retention: Exploring the Role of Machine Learning in E-Commerce. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(2), 26-34.

Kalla, D. (2022). AI-Powered Driver Behavior Analysis and Accident Prevention Systems for Advanced Driver Assistance. *International Journal of Scientific Research and Modern Technology (IJSRMT) Volume*, *1*.

Kalla, D., Kuraku, D. S., & Samaah, F. (2021). Enhancing cyber security by predicting malwares using supervised machine learning models. *International Journal of Computing and Artificial Intelligence*, *2*(2), 55-62.

Kalla, D., Smith, N., Samaah, F., & Polimetla, K. (2021). Facial Emotion and Sentiment Detection Using Convolutional Neural Network. *Indian Journal of Artificial Intelligence Research (INDJAIR)*, *1*(1), 1-13.

Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., & Polam, R. M. (2021). Advanced Machine Learning Models for Detecting and Classifying Financial Fraud in Big Data-Driven. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(3), 39-46.

Katari, A., & Kalla, D. (2021). Cost Optimization in Cloud-Based Financial Data Lakes: Techniques and Case Studies. *ESP Journal of Engineering & Technology Advancements (ESP-JETA)*, *1*(1), 150-157.

Krutthika H. K. & A.R. Aswatha (2020). Design of efficient FSM-based 3D network-on-chip architecture. *International Journal of Engineering Trends and Technology*, *68*(10), 67–73. https://doi.org/10.14445/22315381/IJETT-V68I10P212

Krutthika H. K. & A.R. Aswatha. (2020). FPGA-based design and architecture of network-on-chip router for efficient data propagation. *IIOAB Journal*, *11*(S2), 7–25.

Krutthika H. K. & A.R. Aswatha. (2021). Implementation and analysis of congestion prevention and fault tolerance in network on chip. *Journal of Tianjin University Science and Technology*, *54*(11), 213–231. https://doi.org/10.5281/zenodo.5746712

Krutthika H. K. & Rajashekhara R. (2019). Network-on-chip: A survey on router design and algorithms. *International Journal of Recent Technology and Engineering*, *7*(6), 1687–1691. https://doi.org/10.35940/ijrte.F2131.037619

Krutthika, H. K. (2019, October). Modeling of Data Delivery Modes of Next Generation SOC-NOC Router. In *2019 Global Conference for Advancement in Technology (GCAT)* (pp. 1-6). IEEE.

Nair, T. R., & Krutthika, H. K. (2010). An Architectural Approach for Decoding and Distributing Functions in FPUs in a Functional Processor System. *arXiv preprint arXiv:1001.3781*.

Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

Nandiraju, S. K. K., Chundru, S. K., Vangala, S. R., Polam, R. M., Kamarthapu, B., & Kakani, A. B. (2022). Advance of AI-Based Predictive Models for Diagnosis of Alzheimer's Disease (AD) in Healthcare. *Journal of Artificial Intelligence and Big Data*, *2*(1), 141–152.DOI: 10.31586/jaibd.2022.1340

Narra, B., Vattikonda, N., Gupta, A. K., Buddula, D. V. K. R., Patchipulusu, H. H. S., & Polu, A. R. (2022). Revolutionizing Marketing Analytics: A Data-Driven Machine Learning Framework for Churn Prediction. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *3*(2), 112-121.

Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., & Gangineni, V. N. (2021). An Advanced Machine Learning Models Design for Fraud Identification in Healthcare Insurance. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(1), 26-34.

Penmetsa, M., Bhumireddy, J. R., Chalasani, R., Tyagadurgam, M. S. V., Gangineni, V. N., & Pabbineedi, S. (2021). Next-Generation Cybersecurity: The Role of AI and Quantum Computing in Threat Detection. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(4), 54-61.

Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Data Security in Cloud Computing: Encryption, Zero Trust, and Homomorphic Encryption. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(3), 70-80.

Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., Chundru, S. K., & Vangala, S. R. (2021). Big Text Data Analysis for Sentiment Classification in Product Reviews Using Advanced Large Language Models. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 55-65.

Polu, A. R., Buddula, D. V. K. R., Narra, B., Gupta, A., Vattikonda, N., & Patchipulusu, H. (2021). Evolution of AI in Software Development and Cybersecurity: Unifying Automation, Innovation, and Protection in the Digital Age. *Available at SSRN 5266517*.

Polu, A. R., Narra, B., Buddula, D. V. K. R., Patchipulusu, H. H. S., Vattikonda, N., & Gupta, A. K. BLOCKCHAIN TECHNOLOGY AS A TOOL FOR CYBERSECURITY: STRENGTHS, WEAKNESSES, AND POTENTIAL APPLICATIONS.

Polu, A. R., Vattikonda, N., Gupta, A., Patchipulusu, H., Buddula, D. V. K. R., & Narra, B. (2021). Enhancing Marketing Analytics in Online Retailing through Machine Learning Classification Techniques. *Available at SSRN 5297803*.

Rajiv, C., Mukund Sai, V. T., Venkataswamy Naidu, G., Sriram, P., & Mitra, P. (2022). Leveraging Big Datasets for Machine Learning-Based Anomaly Detection in Cybersecurity Network Traffic. *J Contemp Edu Theo Artific Intel: JCETAI/102*.

Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2022). Designing an Intelligent Cybersecurity Intrusion Identify Framework Using Advanced Machine Learning Models in Cloud Computing. *Universal Library of Engineering Technology*, (Issue).

Tyagadurgam, M. S. V., Gangineni, V. N., Pabbineedi, S., Penmetsa, M., Bhumireddy, J. R., & Chalasani, R. (2021). Enhancing IoT (Internet of Things) Security Through Intelligent Intrusion Detection Using ML Models. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 27-36.

Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2022). Leveraging Artificial Intelligence Algorithms for Risk Prediction in Life Insurance Service Industry. *Available at SSRN 5459694*.

Vangala, S. R., Polam, R. M., Kamarthapu, B., Kakani, A. B., Nandiraju, S. K. K., & Chundru, S. K. (2021). Smart Healthcare: Machine Learning-Based Classification of Epileptic Seizure Disease Using EEG Signal Analysis. *International Journal of Emerging Research in Engineering and Technology*, *2*(3), 61-70.

Vattikonda, N., Gupta, A. K., Polu, A. R., Narra, B., Buddula, D. V. K. R., & Patchipulusu, H. H. S. (2022). Blockchain Technology in Supply Chain and Logistics: A Comprehensive Review of Applications, Challenges, and Innovations. *International Journal of Emerging Research in Engineering and Technology*, *3*(3), 99-107.

License