JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Anemia Prediction Using Machine Learning and NLP

S.Lakshmana Suneel [1], M. Tarani.[2]

[1] Assistant Professor [2] M.Tech Final Semester, Sanketika Vidya Parishad Engineering College, Vishakhapatnam, Andhra Pradesh, India.

slakshmansuneel@gmail.com, taranimamidi7@gmail.com

Abstract: Anemia is a common health disorder caused by a deficiency of hemoglobin or red blood cells, leading to fatigue, weakness, and serious complications if not detected early. Early diagnosis is essential for effective treatment and prevention. This project presents a machine learning-based system designed to predict anemia using important patient health parameters such as hemoglobin levels, red blood cell count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). The dataset is carefully preprocessed to handle missing values and improve overall data quality, which helps enhance the accuracy of the prediction models. Various machine learning algorithms, including Random Forest, Logistic Regression, and Decision Tree, are applied to analyze the clinical data and classify whether an individual is anemic or not. These models are evaluated using performance metrics such as accuracy and precision, with Random Forest providing better results due to its ability to handle complex relationships among features. The system is further integrated into a user-friendly web interface that allows users to input medical data and receive instant predictions. Overall, this project demonstrates the significant role of artificial intelligence in healthcare by enabling early detection of anemia, reducing manual effort, and supporting medical professionals in making accurate and timely decisions

Keywords: RandomForestClassifier, Logistic Regression, Decision tree, XGBoost classifier, LSTM (Long Short -Term Memory), Support Vector Machines (SVM), MinMaxScaler, LabelEncoder

I.Introduction

Anemia is a common health condition caused by a deficiency of hemoglobin or red blood cells, leading to reduced oxygen supply to body tissues and resulting in symptoms like fatigue and weakness; if left undetected, it can lead to serious complications, making early diagnosis essential for effective treatment. This project aims to develop a machine learning-based system for the early prediction of anemia using important clinical parameters such as hemoglobin levels, red blood cell (RBC) count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). The dataset is carefully preprocessed to handle missing values and improve data quality, ensuring accurate and reliable predictions. Various machine learning algorithms, including Random Forest, Logistic Regression, and Decision Tree, are applied and evaluated using performance metrics like accuracy and precision, with Random Forest providing better results due to its ability to handle complex feature relationships. The system is further integrated into a user-friendly web interface that allows users to input medical data and receive instant predictions, demonstrating the potential of artificial intelligence in enhancing healthcare through early detection, reduced manual effort, and improved decision-making support for medical professionals.

II.LITERATURE SURVEY

The literature on anemia detection highlights the growing use of machine learning techniques to improve early diagnosis and clinical decision-making. Traditional methods rely on laboratory tests and manual interpretation, which can be time-consuming and prone to human error. Recent studies have explored data-driven approaches using algorithms such as Logistic Regression, Decision Trees, Support Vector Machines, and Random Forest to analyse patient health parameters like hemoglobin, red blood cell count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH). Researchers have found that ensemble methods, particularly Random Forest, often achieve higher accuracy due to their ability to handle complex and non-linear relationships in medical data. Several works also emphasize the importance of data pre-processing, including handling missing values, normalization, and feature selection, to improve model performance. Additionally, some studies have integrated these predictive models into web-based or clinical decision support systems, enabling real-time anemia detection and assisting healthcare professionals. [9] Despite these advancements, challenges such as limited dataset availability, data imbalance, and model generalization remain areas of ongoing research. Overall, the existing literature demonstrates that machine learning has significant potential to enhance anemia diagnosis by providing faster, more accurate, and scalable solutions compared to traditional approaches.

III.CHALLENGES

One of the major challenges in anemia prediction using machine learning is the availability and quality of medical data, as datasets are often limited, incomplete, or contain missing and inconsistent values that can affect model performance. Another key issue is data imbalance, where the number of anemic and non-anemic cases may not be evenly distributed, leading to biased predictions. Feature selection is also challenging, as identifying the most relevant clinical parameters without introducing redundancy is crucial for achieving accurate results. Additionally, different machine learning models may produce varying outcomes, making it difficult to select the most reliable algorithm without extensive evaluation. There are also concerns related to model generalization, as a model trained on one dataset may not perform well on data from different populations or regions. Integration into real-world healthcare systems presents further difficulties, including ensuring user-friendly interfaces, maintaining patient data privacy, and meeting medical standards. [8] Overall, these challenges highlight the need for robust data preprocessing, careful model selection, and continuous validation to develop an effective and reliable anemia prediction system.uncertainties of human behaviour and climate projections, which can hinder the accuracy of demand forecasting. Furthermore, many organizations find it difficult to transition from traditional physical servers to scalable cloud technology, often lacking the in-house expertise required to manage the complex AI and machine learning systems necessary for real-time, automated decision-making.

IV.PROPOSED METHODOLOGY

The proposed methodology involves collecting a dataset with key health parameters such as hemoglobin, RBC count, MCV, and MCH, followed by data preprocessing to handle missing values and improve quality. The processed data is then split into training and testing sets, and multiple machine learning algorithms like Logistic Regression, Decision Tree, and Random Forest are applied. [7] These models are evaluated using metrics such as accuracy and precision, with Random Forest selected as the best-performing model. Finally, the chosen model is integrated into a user-friendly web application to provide instant anemia predictions based on user input.

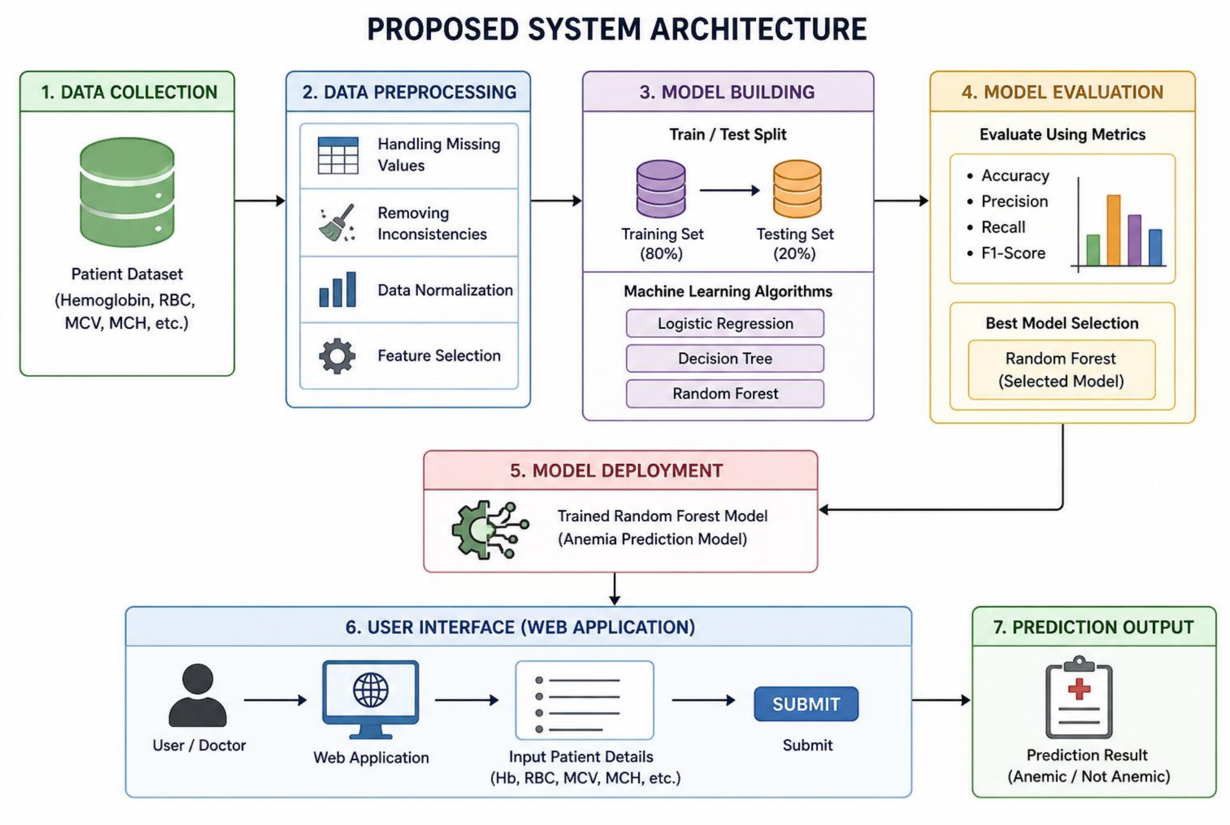

Figure 1: Flow chart of the proposed methodology

V.ALGORITHMS AND TECHNIQUES

The project utilizes a combination of machine learning models and supporting techniques to accurately predict anemia and deliver results through a web-based system:

• **Random Forest Classifier:** An ensemble learning algorithm that builds multiple decision trees and combines their outputs to improve prediction accuracy. It effectively handles complex and non-linear relationships tween features such as hemoglobin, RBC, MCV, and MCH, making it the best-performing model in this project.
• **Logistic Regression:** A statistical classification algorithm used as a baseline model for binary prediction nemic

/ Not Anemic). It is simple, fast, and provides good interpretability of feature importance.
• **Decision Tree Classifier:** A tree-based model that splits data based on feature values, helping to capture decision rules and making the prediction process easy to understand and visualize.
• **Flask Web Framework:** Used to develop the backend of the web application, enabling interaction between the user interface and the trained machine learning model for real-time predictions.
• **Data Pre-processing Techniques:**

o **Handling Missing Values:** Ensures incomplete data does not affect model performance by filling or removing null values.

o **Feature Scaling (Standardization/Normalization):** Applied to bring all input features to a similar range, improving model efficiency and accuracy.

o **Train-Test Split:** Divides the dataset into training and testing sets to evaluate model performance effectively.
• **Model Evaluation Metrics:**

o **Accuracy, Precision, Recall, and F1-Score:** Used to measure and compare the performance of different models and select the most reliable one.

These algorithms and techniques work together to build an efficient, accurate, and user-friendly anemia prediction system. [5]

VI.ARCHITECTURE

The architecture of the anemia prediction system is designed as a multi-layered framework that integrates data processing, machine learning, and user interaction components. It begins with the data layer, where patient medical data such as hemoglobin, RBC, MCV, and MCH is collected and stored. This data is passed to the preprocessing layer, where missing values are handled, features are normalized, and relevant attributes are selected to ensure high-quality input. [2] The processed data is then fed into the model layer, where machine learning algorithms like Random Forest, Logistic Regression, and Decision Tree are trained and used for prediction. The best-performing model, Random Forest, is selected and deployed. Above this, the application layer is built using the Flask web framework, which connects the frontend user interface with the backend prediction model. Users or doctors can input patient details through the web interface, and the system processes this input and sends it to the trained model. Finally, the output layer displays the prediction result (Anemic or Not Anemic) in an easy-to-understand format. This structured architecture ensures efficient data flow, accurate predictions, and a seamless user experience.

VII.OUTPUTS

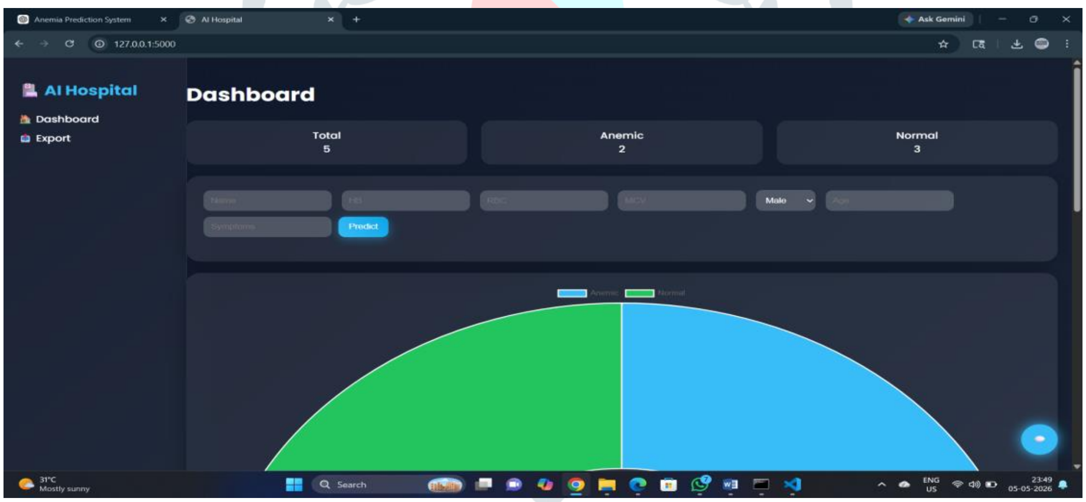

FIG1: HOME SCREEN

The dashboard of the anemia prediction system is designed to provide a clear, interactive, and user-friendly interface for both users and healthcare professionals. It serves as the main control panel where users can easily navigate through different sections such as patient data entry, prediction results, and analytics. The dashboard includes input fields for key medical parameters like hemoglobin (Hb), red blood cell (RBC) count, mean corpuscular volume (MCV), and mean corpuscular hemoglobin (MCH), allowing users to enter patient details efficiently. Once the data is submitted, the system processes the input and displays the prediction result (Anemic or Not Anemic) instantly in a visually appealing format. Additionally, the dashboard may include graphical representations such as charts and reports to help users understand trends and model performance. Built using modern web technologies like HTML, CSS, JavaScript, and integrated with a Flask backend, the dashboard ensures smooth interaction, fast response, and an enhanced user experience, making the system both practical and effective for real-world use

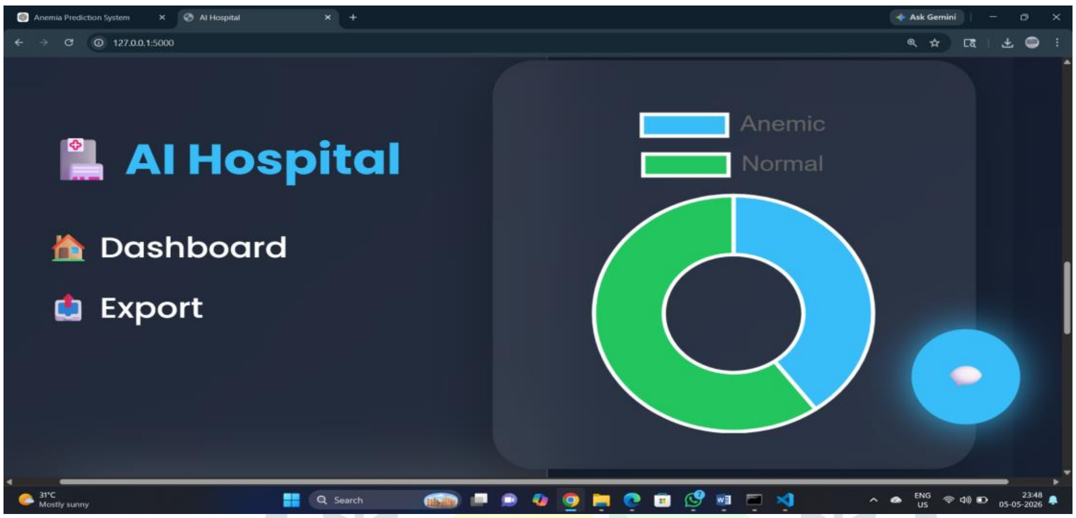

FIG 2: SAMPLE PREDICTION GRAPH

The sample prediction graph displayed on the dashboard provides a clear visual representation of the anemia classification results. It is designed as a donut chart that compares the proportion of “Anemic” and “Normal” cases based on the processed data. Each segment is color-coded for easy understanding, where one colour represents anemic cases and the other represents normal cases, along with a legend for clarity. This type of visualization helps users quickly interpret the model’s output without needing technical knowledge. The graph enhances the overall user experience by making the results more intuitive and visually appealing, allowing doctors or users to analyse patterns and distributions at a glance. Integrated into the dashboard, this graphical representation supports better decision-making by presenting prediction outcomes in a simple, interactive, and informative manner.

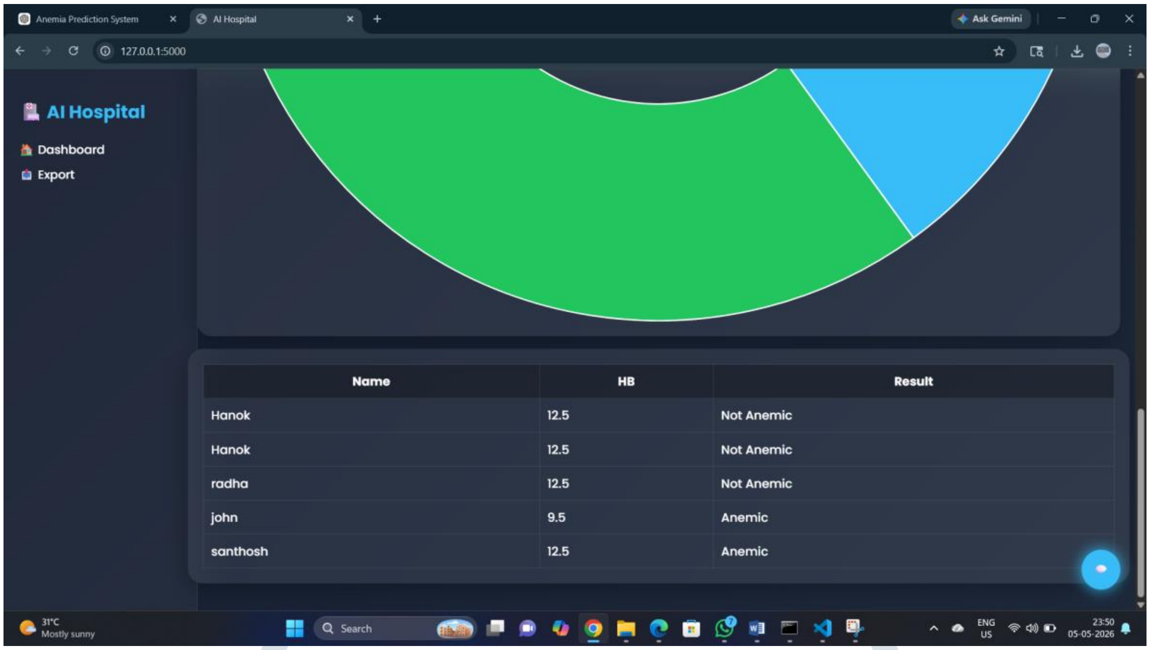

FIG3: PREVIOUS PATIENT DATA

The dashboard also includes a tabular representation of prediction results, which provides a structured and detailed view of patient data alongside the model's output. This table displays key information such as patient name, hemoglobin (HB) values, and the corresponding prediction result (Anemic or Not Anemic). It allows users to easily compare multiple patient records in a single view, making it useful for tracking and analysis. The table is designed with a clean and responsive layout, ensuring readability and ease of use. By combining both graphical visualization (donut chart) and tabular data, the dashboard offers a comprehensive overview of results, helping users and healthcare professionals make quick and informed decisions. Additionally, this feature supports data management and can be extended for functionalities like exporting reports or maintaining patient history.

FIG4: NEW PATIENT DATA INPUT

The dashboard interface of the anemia prediction system is designed to provide a comprehensive and interactive environment for users to analyse and manage patient data efficiently. At the top, summary cards display key statistics such as the total number of records, number of anemic cases, and normal cases, giving a quick overview of the dataset. Below this, an input form allows users to enter patient details including name, hemoglobin level, MCV, MCH, gender, age, and symptoms, enabling real-time prediction through a "Predict" button. The system processes this input using the trained machine learning model and updates the results instantly.

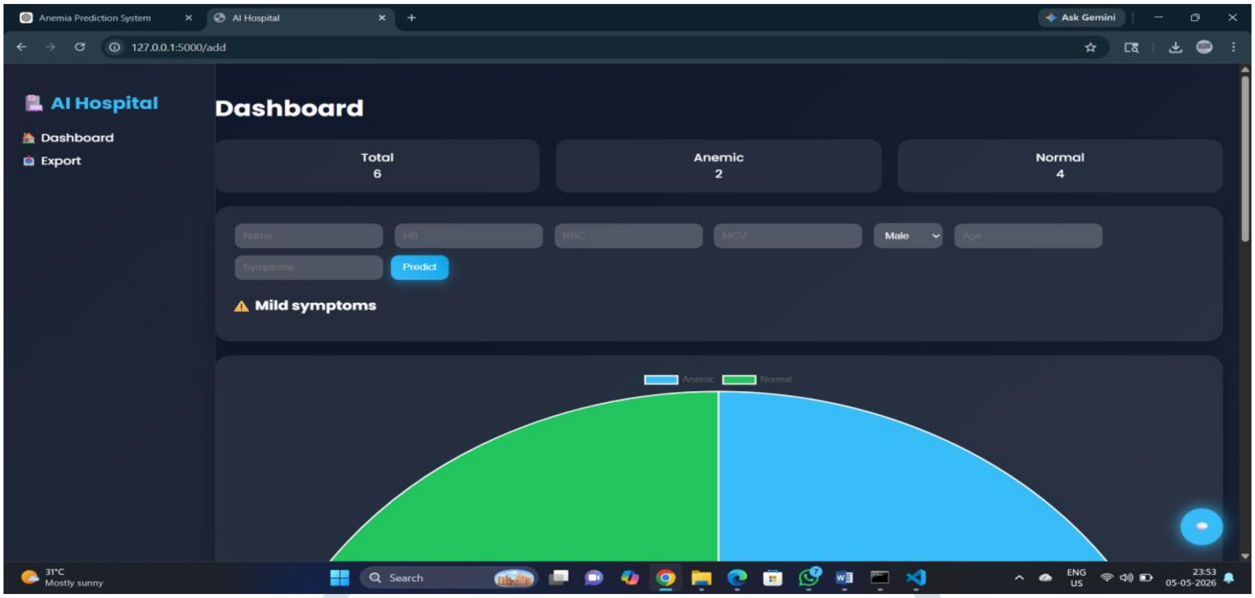

FIG5: FINAL OUTPUT

CONCLUSION:

The anemia prediction system demonstrates how machine learning can be effectively applied in the healthcare domain to support early diagnosis and decision-making. By utilizing key clinical parameters such as hemoglobin, RBC count, MCV, and MCH, along with algorithms like Random Forest, Logistic Regression, and Decision Tree, the system is able to provide accurate and reliable predictions. The integration of these models into a user-friendly web-based dashboard enhances accessibility and usability, allowing users to input data and receive instant results along with visual insights. This project not only reduces manual effort and time required for diagnosis but also improves the overall efficiency of healthcare analysis. Despite certain challenges such as data limitations and model generalization, the system highlights the potential of artificial intelligence in delivering fast, scalable, and cost-effective solutions for medical applications.

REFERENCES

[1] S. Rajaraman, S. Candemir, and G. Thoma, "Visualizing and explaining deep learning predictions for pneumonia detection in chest radiographs," *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, pp. 1–9, 2018.

https://openaccess.thecvf.com/

[2] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, 2017.

https://doi.org/10.1038/nature21056

[3] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770–778, 2016.

https://doi.org/10.1109/CVPR.2016.90

[4] O. Russakovsky et al., "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.

https://doi.org/10.1007/s11263-015-0816-y

[5] D. Shen, G. Wu, and H. Suk, "Deep learning in medical image analysis," *Annu. Rev. Biomed. Eng.*, vol. 19, pp. 221–248, 2017.

https://doi.org/10.1146/annurev-bioeng-071516-044442

[6] J. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.

https://doi.org/10.1016/j.media.2017.07.005

[7] M. Abadi et al., "TensorFlow: Large-scale machine learning on heterogeneous systems," 2016. [Online]. Available:
https://www.tensorflow.org/

[8] F. Chollet, "Keras: Deep learning library for Python," 2015. [Online]. Available:

https://keras.io/

[9] G. Huang, Z. Liu, L. van der Maaten, and K. Weinberger, "Densely connected convolutional networks," *Proc. IEEE CVPR*, pp. 4700–4708, 2017.

https://doi.org/10.1109/CVPR.2017.243

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.

https://doi.org/10.1145/3065386

[11] S. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

https://arxiv.org/abs/1409.1556

[12] C. Szegedy et al., "Going deeper with convolutions," *Proc. IEEE CVPR*, pp. 1–9, 2015.

https://doi.org/10.1109/CVPR.2015.7298594

[13] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. Dudley, "Deep learning for healthcare: Review, opportunities and challenges," *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, 2018.

https://doi.org/10.1093/bib/bbx044

[14] T. Rahman et al., "Exploring the effect of image enhancement techniques on COVID-19 detection using chest X-ray images," *Comput. Biol. Med.*, vol. 132, p. 104319, 2021.

https://doi.org/10.1016/j.compbiomed.2021.104319

[15] P. Kaur, G. Singh, and P. Kaur, "Intelligent diagnosis of anemia using machine learning techniques," *Proc. Int. Conf. Comput. Intell. Data Sci.*, pp. 1–6, 2020.

[16] S. R. Dubey, S. K. Singh, and R. K. Singh, "Multimodal deep learning for medical diagnosis: A review," *IEEE Access*, vol. 9, pp. 121001–121020, 2021.https://ieeexplore.ieee.org/

[17] H. Greenspan, B. van Ginneken, and R. Summers, "Guest editorial: Deep learning in medical imaging: Overview and future promise," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016. https://doi.org/10.1109/TMI.2016.2535868

[18] J. Arevalo, F. González, R. Ramos-Pollán, J. Oliveira, and M. Guevara Lopez, "Representation learning for mammography mass lesion classification with deep learning," *Comput. Methods Programs Biomed.*, vol. 127, pp. 248–257, 2016.

https://doi.org/10.1016/j.cmpb.2016.03.014

[19] M. Anthimopoulos et al., "Lung pattern classification for interstitial lung diseases using a deep CNN," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1207–1216, 2016.

https://doi.org/10.1109/TMI.2016.2535865

BIBLOGRAPHY

S. Lakshmana Suneel is currently working as an Assistant Professor at Sanketika Vidya Parishad Engineering College (SVPEC) and has 1 year of teaching experience. Prior to joining academia, he worked in the software industry in SAP, gaining valuable industry experience. He holds an M.Tech in Computer Science and has a strong interest in Java, Python, Database Management Systems (DBMS), and Software Engineering. His combination of academic knowledge and industry experience enables him to provide students with both theoretical concepts and practical, real-world perspectives.

M. Tarani is currently pursuing her final semester of M. at Sanketika Vidya Parishad Engineering College, which is accredited with an 'A' grade by NAAC, affiliated to Andhra University, and approved by AICTE. With a keen interest in Machine Learning and Artificial Intelligence, She has undertaken her postgraduate project titled "Anemia prediction using machine learning and NLP" The project has been successfully carried out under the guidance of S. Lakshmana Suneel is currently working as an Assistant Professor, SVPEC.