IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Coronary Heart Disease Prediction: Optimized Light GBM Model For Superior Performance

Mrs. Kommuri Vijitha[1], Ramella Hima Bindu[2], Inteti Lahari[3], Chitturi Adinarayana[4]

#1 Assistant Professor in the Department of IT, NRI INSTITUTE OF TECHNOLOGY, AGIRIPALLI.

#2#3#4 B.Tech with Specialization of Information Technology in NRI INSTITUTE OF TECHNOLOGY, AGIRIPALLI.

Abstract: Unfortunately, there is currently no cure for coronary heart disease (CHD), a deadly cardiac condition. Identifying coronary artery disease accurately and quickly is crucial for effective patient therapy. Treatments can be initiated sooner and patient outcomes can be improved with early detection. Using an optimised LightGBM classifier, the "HY_OptGBM" model forecasts CHD. When it comes to predictive modelling, the gradient boosting framework LightGBM is both efficient and accurate. Optimisation of the LightGBM classifier is achieved by adjustments to the hyperparameters and loss function. The accuracy and efficiency of model training are both enhanced by this optimisation strategy. Data on coronary heart disease from the Framingham Heart Institute is used to evaluate the model's performance. The algorithm reliably predicts CHD using this data, which paves the way for early detection and, maybe, reduced treatment expenses. Additionally, it presents a Voting Classifier (RF + AdaBoost) that can identify Coronary Heart Disease with a 99% accuracy rate. This ensemble model using Random Forest and AdaBoost does a good job of differentiating CHD patterns. User registration and sign-in are made easier with the use of an intuitive Flask framework that integrates with SQLite, allowing for easier usability testing. This streamlined interface may make the use of machine learning technologies easier for the stakeholders involved in CHD identification.

Index terms *- OPTUNA, ML, CHD, hyperparameter optimisation, LightGBM, loss function.*

1. INTRODUCTION

Coronary heart disease (CHD) is a prevalent cardiovascular illness caused by atherosclerotic plaques in the coronary arteries, which restrict blood flow to the heart muscle. Chest pain, angina, dyspnoea, palpitations, and heart failure are some of the symptoms. The heart muscle can be irreparably damaged and quality of life reduced in the event of a heart attack caused by coronary heart disease. Important steps should be taken to identify and control CHD, including medical intervention and behavioural modifications [1].

Cure rates and treatment costs are both improved by early detection of CHD. Several data mining methods and machine learning algorithms have found extensive use in the medical field as a result of recent advancements in the field and falling data storage costs [2, 3, 4, 5, 6]. Medical research, drug discovery, biomedicine, and supplemental diagnostics all rely on data mining tools. By utilising data

mining technology, we can glean hidden disease information from massive volumes of unstructured medical data, build models to forecast diseases, and assess the results.

Providing affordable, high-quality therapy is not without its challenges for healthcare professionals. In order to avoid squandering funds, hospitals rely on doctors who are well-versed in their fields and can accurately diagnose patients. In medical settings, data mining technology is both useful and essential. Any classification method's performance is significantly affected by the optimal hyperparameters [7], [8]. The accuracy of a classification algorithm can be enhanced by selecting optimal hyperparameters. The hyperparameters of the LightGBM model were optimised in this work using OPTUNA [9]. As a result, our research focused on finding the optimal hyperparameters among the available options. Hyperparameters can be optimised using grid and random searches. Optuna hyperparametric search is an additional method. Due to the fact that LightGBM's performance is hyperparameter-dependent, conventional random and grid search approaches are inefficient and wasteful as they do not take into account previous optimisations. OPTUNA adjusts hyperparameters based on what it learns from previous optimisations. For optimising hyperparameters in this study, OPTUNA was selected.

The accuracy of the model is affected by the loss function [10]. A loss function that incorporates category weight α and a component that modifies sample difficulty weight γ, based on cross-entropy loss, is suggested in this study. The unequal distribution of positive and negative samples was the focus of this investigation. Model performance may be enhanced by utilising the focus loss feature as well. The focused loss function was employed in this work to enhance the default loss function of the LightGBM [11] model, which is utilised to predict CHD.

2. LITERATURE SURVEY

a) Obesity and Coronary Heart Disease: Epidemiology, Pathology, and Coronary Artery Imaging

Obesity and Coronary Heart Disease: Epidemiology, Pathology, and Coronary Artery Imaging - ScienceDirect

Overweight and obesity contribute to the development of cardiovascular disease (CVD) in general and coronary heart disease (CHD) in particular in part by their association with traditional and nontraditional CVD risk factors. Obesity is also considered to be an independent risk factor for CVD. The metabolic syndrome, of which central obesity is an important component, is strongly associated with CVD including CHD. There is abundant epidemiologic evidence of an association between both overweight and obesity and CHD. Evidence from postmortem studies and studies involving coronary artery imaging is less persuasive. Recent studies suggest the presence of an obesity paradox with respect to mortality in persons with established CHD. Physical activity and preserved cardiorespiratory fitness attenuate the adverse effects of obesity on CVD events. Information concerning the effect of intentional weight loss on CVD outcomes in overweight and obese persons is limited.

b) eDoctor: machine learning and the future of medicine

https://pubmed.ncbi.nlm.nih.gov/30102808/

Machine learning (ML) is a burgeoning field of medicine with huge resources being applied to fuse computer science and statistics to medical problems. Proponents of ML extol its ability to deal with large, complex and disparate data, often found within medicine and feel that ML is the future for biomedical research, personalized medicine, computer-aided diagnosis to significantly advance global health care. However, the concepts of ML are unfamiliar to many medical professionals and there is untapped potential in the use of ML as a research tool. In this article, we provide an overview of the theory behind ML, explore the common ML algorithms used in medicine including their pitfalls and discuss the potential future of ML in medicine.

c) Artificial Intelligence in Drug Treatment

http://pubmed.ncbi.nlm.nih.gov/31348869/

The most common applications of artificial intelligence (AI) in drug treatment have to do with matching patients to their optimal drug or combination of drugs, predicting drug-target or drug-drug interactions, and optimizing treatment protocols. This review outlines some of the recently developed AI methods aiding the drug treatment and administration process. Selection of the best drug(s) for a patient typically requires the integration of patient data, such as genetics or proteomics, with drug data, like compound chemical descriptors, to score the therapeutic efficacy of drugs. The prediction of drug interactions often relies on similarity metrics, assuming that drugs with similar structures or targets will have comparable behavior or may interfere with each other. Optimizing the dosage schedule for administration of drugs is performed using mathematical models to interpret pharmacokinetic and pharmacodynamic data. The recently developed and powerful models for each of these tasks are addressed, explained, and analyzed here.

d) Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification

https://pubmed.ncbi.nlm.nih.gov/32245523/

Background: Machine learning is a sub-field of artificial intelligence, which utilises large data sets to make predictions for future events. Although most algorithms used in machine learning were developed as far back as the 1950s, the advent of big data in combination with dramatically increased computing power has spurred renewed interest in this technology over the last two decades.

Main body: Within the medical field, machine learning is promising in the development of assistive clinical tools for detection of e.g. cancers and prediction of disease. Recent advances in deep learning technologies, a sub-discipline of machine learning that requires less user input but more data and processing power, has provided even greater promise in assisting physicians to achieve accurate diagnoses. Within the fields of genetics and its sub-field epigenetics, both prime examples of complex data, machine learning methods are on the rise, as the field of personalised medicine is aiming for treatment of the individual based on their genetic and epigenetic profiles.

Conclusion: We now have an ever-growing number of reported epigenetic alterations in disease, and this offers a chance to increase sensitivity and specificity of future diagnostics and therapies. Currently, there are limited studies using machine learning applied to epigenetics. They pertain to a wide variety of disease states and have used mostly supervised machine learning methods.

e) Machine learning for cardiology

https://pubmed.ncbi.nlm.nih.gov/34338485/

This paper reviews recent cardiology literature and reports how artificial intelligence tools (specifically, machine learning techniques) are being used by physicians in the field. Each technique is introduced with enough details to allow the understanding of how it works and its intent, but without delving into details that do not add immediate benefits and require expertise in the field. We specifically focus on the principal Machine learning based risk scores used in cardiovascular research. After introducing them and summarizing their assumptions and biases, we discuss their merits and shortcomings. We report on how frequently they are adopted in the field and suggest why this is the case based on our expertise in machine learning. We complete the analysis by reviewing how corresponding statistical approaches compare with them. Finally, we discuss the main open issues in applying machine learning tools to cardiology tasks, also drafting possible future directions. Despite the growing interest in these tools, we argue that there are many still underutilized techniques: while neural networks are slowly being incorporated in cardiovascular research, other important techniques such as semi-supervised learning and federated learning are still underutilized. The former would allow practitioners to harness the information contained in large datasets that are only partially labeled, while the latter would foster collaboration between institutions allowing building larger and better models.

3. METHODOLOGY

i) Proposed Work:

The proposed system enhances a LightGBM model for CHD prediction by optimising it, testing its efficacy, applying ensemble techniques, receiving user input, and including an authentication and user-friendly interface. Optimisation and ensemble approaches improve the accuracy of coronary heart disease prediction. To maximise the precision of its predictions, LightGBM fine-tunes its parameters and loss functions. Several areas of healthcare can benefit from the system due to its adaptability and relevance [11,26]. Additionally, it presents a Voting Classifier (RF + AdaBoost) that can identify Coronary Heart Disease with a 99% accuracy rate. This ensemble model using Random Forest and AdaBoost does a good job of differentiating CHD patterns. User registration and sign-in are made easier with the use of an intuitive Flask framework that integrates with SQLite, allowing for easier usability testing. Stakeholders in CHD detection may find it easier to use machine learning algorithms with this streamlined interface [2, 3, 4, 5, 6].

ii) System Architecture:

Machine learning models, particularly those dealing with massive training and datasets, benefit from simpler configurations. With all those features, OPTUNA is a fantastic framework for hyperparametric optimisation. Figure 1 displays the improved LightGBM model architecture. In Fig. 1, we can observe that during the search, every worker executed the target function.

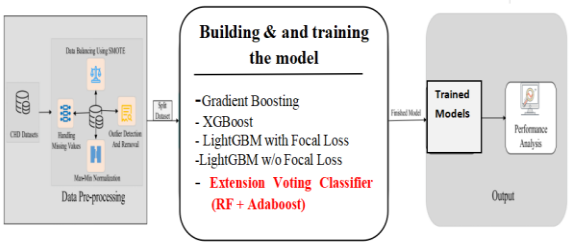

Fig 1 Proposed architecture

iii) Dataset collection:

It is necessary to load and analyse Framingham Heart Disease data in order to ascertain its structure, attributes, and content. Exploring potential risk factors for cardiovascular disease is the goal of the Framingham Heart Study (FHS). There were 5,209 men and women in Framingham, MA, ranging in age from 30 to 62 in 1948. The following cohorts began: 1971 Offspring, 1994 Omni, 2002 Third Generation, 2004 New Offspring Spouse, and 2003 Second Generation Omni. Cardiovascular and cerebrovascular disease research makes up the bulk of the dataset. The data set includes biological specimens, phenotypic data, samples, images, demographic information, physiological data, vascular functional data, and molecular genetic data for the participants.Teamwork between the National Heart, Lung, and Blood Institute and Boston University.

	Sex	Age	Education	CurrentSmoker	CigsPerDay	BPMeds	PrevalentStroke	PrevalentHyp
0	1	39	1	0	0.0	0.0	0	0
1	0	46	0	0	0.0	0.0	0	0
2	1	48	0	1	20.0	0.0	0	0
3	0	61	1	1	30.0	0.0	0	1
4	0	46	1	1	23.0	0.0	0	0

Fig 2 Framingham Heart Disease Data

iv) Data Processing:

Data processing transforms unstructured data into actionable insights for businesses. Professionals in the field of data science collect, sort, clean, validate, analyse, and present their findings in written or visual formats. Manual, mechanical, or electronic data processing is possible. Data should be more useful, and choices should be less complicated. That way, companies may improve their operations and make important decisions more quickly. Software engineering and other forms of automated data processing help with this. Quality management and decision-making can benefit from the insights gleaned from big data.

v) Feature selection:

When building a model, feature selection is used to pick the most relevant, consistent, and non-redundant features. Reducing database sizes gradually is critical as database quantity and variety continue to grow. Improving the accuracy of prediction models while decreasing their computational burden is the primary goal of feature selection.

Identifying which features are most important for ML algorithms is a critical aspect of feature engineering. By removing irrelevant or redundant characteristics and keeping just the most crucial ones, feature selection methods help reduce the number of input variables used by machine learning models. There are a number of benefits to selecting features in advance rather than letting the machine learning model decide which ones are most important.

4. EXPERIMENTAL RESULTS

Precision: A high level of accuracy in classifying positive instances or samples is known as precision. Accuracy is determined by applying the following formula:

Precision = True positives/ (True positives + False positives) = TP/(TP + FP)

$$\text{Precision} = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

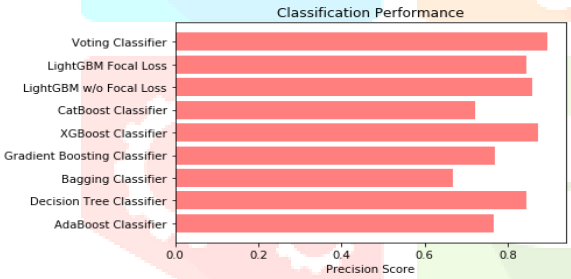

Fig 6 Precision comparison graph

Recall: The ability of a model to identify all pertinent instances of a class is assessed by machine learning recall. By comparing the number of correctly predicted positive observations to the total number of positives, it reveals how well a model captures examples of a class.

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

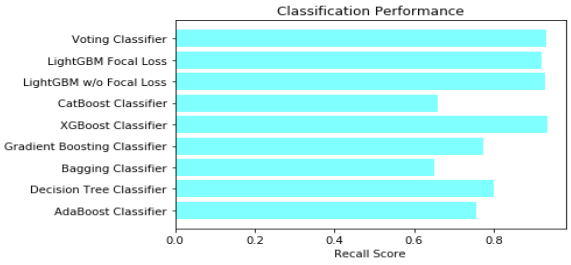

Fig 7 Recall comparison graph

Accuracy: An indicator of a model's performance is the proportion of correct classification predictions.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

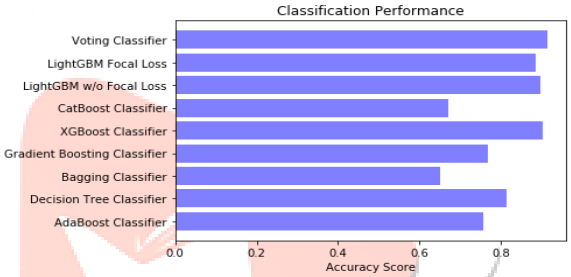

Fig 8 Accuracy graph

F1 Score: For imbalanced datasets, the F1 Score—the harmonic mean of recall and accuracy—is excellent since it balances false positives and negatives.

$$\text{F1 Score} = 2 * \frac{\text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}} * 100$$

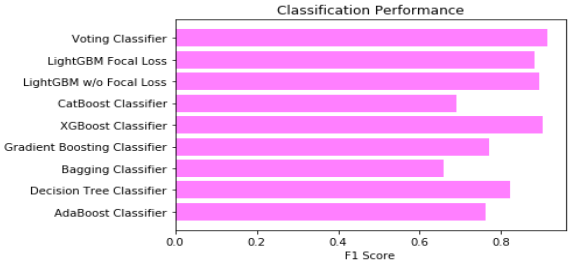

Fig 9 F1Score

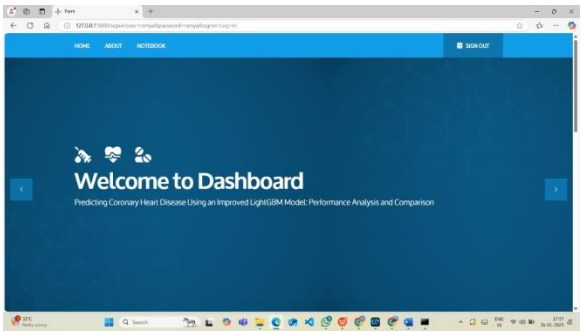

Fig 10 home page

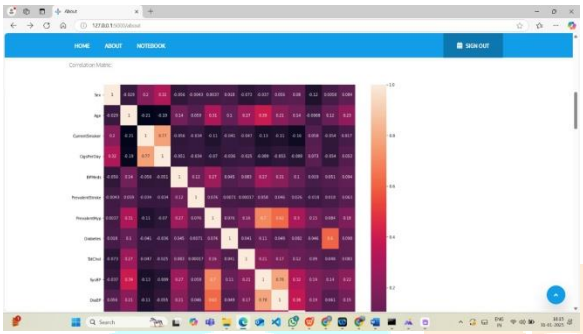

Fig 11 metric table

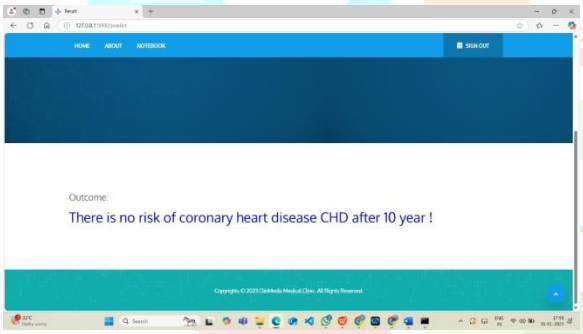

Fig 12 results page

5. CONCLUSION

The HY_OptGBM prediction model successfully indicates CHD by utilising an optimised LightGBM classifier and an enhanced loss function. Measures of the model's predictive abilities include accuracy, precision, recall, and F score. The HY_OptGBM model is enhanced by optimisers using robust loss functions and classifiers. The accuracy of the model's CHD detection and prediction is improved by these adjustments [2, 3, 4, 5, 6]. The accuracy and robustness of a system can be enhanced using an ensemble approach, which integrates predictions from several models. By using a variety of models, advanced ensemble methods like the Voting Classifier are able to improve prediction performance and reach 99% accuracy. System testing is made easier with Flask's user-friendly UI and secure authentication. We assure usability and security by simplifying data entering for system performance evaluation with this interface.

6. FUTURE SCOPE

Additional features or data sources can be added to the HY OptGBM model in future studies to enhance its ability to forecast coronary heart disease. In order to get a whole picture, medical records could be included. It would be beneficial for future research to evaluate the model's robustness and generalisability on larger and more diverse datasets. This will demonstrate the model's ability to adjust to different data distributions. It is possible to assess the HY OptGBM model's performance and superiority in CHD prediction by comparing it to other advanced machine learning algorithms [12,13]. Not only may coronary heart disease but other cardiovascular illnesses or circumstances be predicted using the proposed method. This expansion has the potential to revolutionise cardiology by delivering a versatile tool for prediction.

REFERENCES

[1] N. Katta, T. Loethen, C. J. Lavie, and M. A. Alpert, ‘‘Obesity and coronary heart disease: Epidemiology, pathology, and coronary artery imaging,’’ Current Problems Cardiol., vol. 46, no. 3, Mar. 2021, Art. no. 100655, doi: 10.1016/j.cpcardiol.2020.100655.

[2] G. S. Handelman, H. K. Kok, R. V. Chandra, A. H. Razavi, M. J. Lee, and H. Asadi, ‘‘EDoctor: Machine learning and the future of medicine,’’ J. Internal Med., vol. 284, no. 6, pp. 603–619, Sep. 2018, doi: 10.1111/joim.12822.

[3] E. L. Romm and I. F. Tsigelny, ‘‘Artificial intelligence in drug treatment,’’ Annu. Rev. Pharmacol. Toxicol., vol. 60, no. 1, pp. 353–369, Jan. 2020, doi: 10.1146/annurev-pharmtox-010919-023746.

[4] L. Lo Vercio, K. Amador, J. J. Bannister, S. Crites, A. Gutierrez, M. E. MacDonald, J. Moore, P. Mouches, D. Rajashekar, S. Schimert, N. Subbanna, A. Tuladhar, N. Wang, M. Wilms, A. Winder, and N. D. Forkert, ‘‘Supervised machine learning tools: A tutorial for clinicians,’’ J. Neural Eng., vol. 17, no. 6, Dec. 2020, Art. no. 062001, doi: 10.1088/1741-2552/abbff2.

[5] S. Rauschert, K. Raubenheimer, P. E. Melton, and R. C. Huang, ‘‘Machine learning and clinical epigenetics: A review of challenges for diagnosis and classification,’’ Clin. Epigenetics, vol. 12, no. 1, p. 51, Apr. 2020, doi: 10.1186/s13148-020-00842-4.

[6] Y. Arfat, G. Mittone, R. Esposito, B. Cantalupo, G. M. De Ferrari, and M. Aldinucci, ‘‘Machine learning for cardiology,’’ Minerva Cardiol. Angiol., vol. 70, no. 1, pp. 75–91, Mar. 2022, doi: 10.23736/s2724- 5683.21.05709-4.

[7] S. Nematzadeh, F. Kiani, M. Torkamanian-Afshar, and N. Aydin, ‘‘Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases,’’ Comput. Biol. Chem., vol. 97, Apr. 2022, Art. no. 107619, doi: 10.1016/j.compbiolchem.2021.107619.

[8] M. Liang, B. An, K. Li, L. Du, T. Deng, S. Cao, Y. Du, L. Xu, X. Gao, L. Zhang, J. Li, and H. Gao, ‘‘Improving genomic prediction with machine learning incorporating TPE for hyperparameters optimization,’’ Biology, vol. 11, no. 11, p. 1647, Nov. 2022, doi: 10.3390/ biology11111647.

[9] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, ‘‘OPTUNA: A nextgeneration hyperparameter optimization framework,’’ in Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Anchorage, AK, USA, 2019, pp. 2623–2631.

[10] M. Yeung, E. Sala, C.-B. Schönlieb, and L. Rundo, ‘‘Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation,’’ Computerized Med. Imag. Graph., vol. 95, Jan. 2022, Art. no. 102026, doi: 10.1016/j.compmedimag.2021.102026.

[11] G. Ke et al., ‘‘LightGBM: A highly efficient gradient boosting decision tree,’’ in Proc. 31st Int. Conf. Neural Inf. Process. Syst., Long Beach, CA, USA, 2017, pp. 3149–3157.

[12] O. Goldman, O. Raphaeli, E. Goldman, and M. Leshno, ‘‘Improvement in the prediction of coronary heart disease risk by using artificial neural networks,’’ Qual. Manage. Health Care, vol. 30, no. 4, pp. 244–250, Jul. 2021, doi: 10.1097/qmh.0000000000000309.

[13] Z. Du, Y. Yang, J. Zheng, Q. Li, D. Lin, Y. Li, J. Fan, W. Cheng, X.-H. Chen, and Y. Cai, ‘‘Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation,’’ JMIR Med. Informat., vol. 8, no. 7, Jul. 2020, Art. no. e17257, doi: 10.2196/17257.

[14] J. K. Kim and S. Kang, ‘‘Neural network-based coronary heart disease risk prediction using feature correlation analysis,’’ J. Healthcare Eng., vol. 2017, Sep. 2017, Art. no. 2780501, doi: 10.1155/2017/2780501.

[15] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, and T. Kitai, ‘‘Artificial intelligence in precision cardiovascular medicine,’’ J. Amer. College Cardiol., vol. 69, no. 21, pp. 2657–2664, 2017, doi: 10.1016/j.jacc.2017.03.571.

Author Profile:

Mrs. KOMMURI VIJITHA is currently working as a Assistant Professor in the Department of Information Technology at NRI Institute of Technology(NRITT), Agiripalli, Vijayawada, Andhra Pradesh, India. She had 4 years of academic experience. She carned master's degree in Computer Science and Engineering. She had Published more than 2 research papers in various International Journals Research areas are Artificial Intelligence and Machine Learning.

Email- kommurivijitha1997@gmail.com

I am Ramella Hima Bindu, currently pursuing a B.Tech in Information Technology at NRI Institute of Technology with an 85% grade. My areas of interest include AI, Data Science, and Cloud Computing. I have earned certifications such as NPTEL's "Joy of Computing using Python" and have completed internships in Data Science and DevOps. Through these experiences, I gained hands-on expertise in machine learning, predictive analytics, cloud automation, and CI/CD pipelines. I have worked on projects involving data analysis and automation. With strong problem-solving, communication, and leadership skills, I am eager to contribute to innovative and impactful projects.

I am Inteti Lahari, currently enrolled in a B.Tech program in Information Technology at NRI Institute of Technology, Agiripalli. She is skilled in programming languages such as Python and possesses in-depth knowledge of Data Science. She had successfully completed an internship in Data Science, and gaining practical experience in data analysis, predictive modeling, and visualization techniques. She is passionate about Artificial Intelligence development, with a keen interest in creating innovative and scalable AI-driven solutions to address complex real-world problems.

I am Chitturi Adinarayana, currently pursuing a B.Tech in Information Technology at NRI Institute of Technology. I am deeply interested in Web Development. I have completed certifications in Joy Of Computing using Python Data Structure and Algorithms using Java along with internships in Data science and DevOps. Additionally, I have learned coding experience in Java, python and I am knowledgeable about various Al tools and their applications across different domains. I am passionate about building scalable applications and solving real-world problems through technology.