



International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 4, April 2025

Bigdata Analysis of Pesticide Poisoning in Rural Worker Using Machine Learning Algorithm

**Patnana Chandramouli, Pachipenta Supraja, Penta Indumathi, Kinthali Sarat Kumar,
Marada Lakshmi Prasanna**

UG Scholar, Dept. of Computer Science Engineering (Artificial Intelligence and Data Science), Satya Institute of
Technology and Management, Vizianagaram, India

Seepana Ratna Kumari

Assistant Professor, Dept. of Computer Science Engineering (Artificial Intelligence and Data Science), Satya Institute
of Technology and Management, Vizianagaram, India

ABSTRACT: It is in recent times that the working conditions of the rural and farm workers have drawn special attention on account of escalating exposure to toxins like pesticides. Detection and diagnosis of pesticide poisonings at the initial stages become crucial to save the patient at the earliest stage of medical interventions and to forestall long-term complications. Even so, few rural communities still lack access to high-tech equipment and techniques in diagnostics. To resolve this problem, this project introduces a data-driven methodology using Supervised Machine Learning (SML) methods for the efficient and accurate diagnosis of pesticide poisoning using clinical and biochemical markers. The main contribution of this research is the development and application of a systematic methodology called Data Refinement Cycle with Supervised Machine Learning (DRC-SML). This model was created to offer a complete, applicable, and reproducible methodology that directs a Data Science project from start to finish. It consists of phases like data collection, data preprocessing, missing value handling, encoding and scaling, model selection, performance metrics, and preparation for deployment. In contrast to most current methodologies, which only deal with isolated phases of the data pipeline, DRC-SML guarantees an overall workflow that facilitates a smooth transition between the steps and uniform results.

KEYWORDS: Data Science, Supervised Learning, Data Refinement, Pesticide Poisoning, Rural Healthcare, Machine Learning, Rough Set Theory, Decision Rules, Healthcare Diagnostics

I. INTRODUCTION

The use of data science and machine learning in healthcare has grown significantly in recent years, especially in terms of enhancing decision-making and diagnostic precision. Predictive models can be revolutionary in public health, which is one of the many fields where these methods are showing promise. Pesticide poisoning is one of the major health problems that rural labourers, especially those employed in the agricultural industry, encounter. Serious health effects, such as acute poisoning and persistent long-term disorders, can result from pesticide exposure, whether it occurs by ingesting, skin contact, or inhalation.

Reducing the negative impacts of pesticide exposure requires early diagnosis and prompt action. among order to diagnose pesticide toxicity among rural labourers, this research suggests using a supervised learning model. The algorithm will be taught to categorise and forecast the risk of pesticide poisoning in people who work in rural agricultural settings by utilising a combination of clinical data, environmental exposure measures, and health-related markers. Historical data on workers exposed to pesticides and experiencing different levels of poisoning symptoms will be analysed using supervised learning, more especially classification algorithms. In rural healthcare settings, where resources are frequently scarce and prompt diagnosis may significantly impact patient outcomes, the aim of this research is to develop an accurate and efficient diagnostic tool that can be implemented. By giving medical professionals insightful information and facilitating more focused actions, the use of data science approaches in this situation has the potential to completely transform the diagnosis and treatment of pesticide poisoning. To find the best model for forecasting pesticide toxicity, the research will investigate a range of supervised learning methods, such as

logistic regression, decision trees, random forests, and support vector machines. In order to improve the health and well-being of agricultural workers, the results of this study should benefit the larger field of public health by providing a scalable, data-driven method for identifying occupational disorders in rural areas.

The suggested method uses machine learning techniques, notably Random Forest and Decision Tree algorithms, to improve the detection of pesticide toxicity in rural labourers. In order to anticipate the risk of pesticide poisoning, this data-driven method uses worker demographics, environmental exposure levels, and past health data to train a model. Compared to conventional techniques, the system will be able to diagnose problems more quickly and accurately by analysing intricate patterns and relationships in the data. An ensemble of decision trees that vote on the final classification will be created using Random Forest, which is renowned for its resilience against overfitting and capacity to handle big datasets. On the other hand, decision trees will offer a simple, comprehensible method of comprehending how various elements (such worker health history or pesticide exposure) affect the diagnosis. The suggested approach will provide a strong, scalable way to diagnose pesticide toxicity by integrating both algorithms, especially in rural areas with limited resources.

II. RELATED WORK

The field of data science has greatly expanded beyond its roots in conventional statistics. The exponential rise in the volume of data created and stored worldwide is a startling sign of this progression. Cremin et al. [1] estimate that this volume reached about 44 zettabytes in early 2020, and by 2025, the daily quantity of data created globally is expected to exceed 463 exabytes. Often referred to as "big data," this enormous collection of data is distinguished by its significant volume and diverse range of data kinds.

Even though the subject of data science has grown significantly, there are still many obstacles to overcome before initiatives in this area can be completed successfully. About 87% of Data Science projects do not make it to the production stage, according to data from Saltz and Krasteva [2]. Data science is an interdisciplinary discipline with applications in many different fields. To achieve the desired outcomes, cooperation with subject-matter specialists who possess a thorough understanding of the problems at hand is essential at every stage of the process. Data scientists need to be well-versed in Knowledge Discovery in Databases (KDD), which requires an understanding of databases, machine learning, statistics, and computer science [3].

As a guide for data scientists during the KDD process, data science models frequently adhere to cyclical frameworks called the data lifecycle. Lack of interpretability in complex models and the existence of noisy or low-quality data are frequent problems in this context that can jeopardise the models' efficacy and dependability [4]. Jain and Kushagra [5] have pointed out that the quality of a generated model is inextricably tied to the data that is supplied. Finding pertinent data, merging datasets, cleansing data, producing new data, and gleaning new characteristics from preexisting data are all part of this process. In conclusion, the most time-consuming and perhaps most important phase in a Data Science project's lifetime is data preparation.

According to De Bie et al. [6], machine learning techniques are a key component of a data scientist's toolkit. Over the past 20 years, these techniques—which range from very straightforward ways to intricate strategies like deep learning—have become more and more popular. It is important to note, though, that these approaches frequently presume the availability of large amounts of high-quality data, which in reality creates further difficulties. Supervised, unsupervised, and reinforcement learning are the three primary categories of machine learning.

In supervised learning, a dataset and its corresponding class label explain instances, while experts label the data. In order to enable the machine to categorise fresh unlabelled instances, the main objective is to construct a classifier based on existing examples [7]. A useful mathematical tool for handling a particular kind of uncertainty and imprecision is Rough Set Theory (RST), which was first put out by Pawlak [8] and then examined by Achariva and Abraham [9]. RST, whether used alone or in conjunction with other machine learning models, has proved its usefulness in tackling real-world machine learning issues.

The focus of a data science project is typically data that is embedded in a particular environment. Data science initiatives have helped with a wide range of issues, including marketing, financial analysis, and healthcare analytics [10]. This work tackles a public health issue that hasn't been investigated in data science before. Pesticide poisoning among agricultural labourers is a serious issue that causes substantial social and economic costs globally, according to Peña-Fernández et al. [11]. This is because there aren't enough standardised tests for biological diagnosis, and there aren't enough qualified medical personnel to handle these situations. In order to derive information and value from massive amounts of data, the area of data science is inherently interdisciplinary, incorporating aspects of computer science, mathematics, and statistics [12]. According to a poll by [13] that involved nonprofits and experts in the sector, 85% of participants said that the implementation of more streamlined and reliable procedures might increase the efficacy of Data Science initiatives.

In the field of machine learning, Laturiuw and Singgalen [14] carried out a thorough analysis of algorithms like Support Vector Machines (SVM), Naïve Bayes (NB), Decision Trees (DT), and k-Nearest Neighbours (k-NN) algorithms. They found that SVM and DT performed better in their study when they used SMOTE [15] operators to handle unbalanced data. The article by [16] reviews various Machine Learning algorithms, including DT, SVM, k-NN, and NB, highlighting the superior performance of SVM in terms of accuracy. It should be emphasised that the effectiveness of these methods significantly depends on the domain and specific dataset, which might vary in different settings. Algorithms such as logistic regression, KNN, decision tree, and SVM were assessed on datasets pertaining to breast cancer, heart disease, iris, and wine quality in a thorough machine learning review by Velayutham et al. [17]. Overall, logistic regression was the best-performing algorithm, particularly in the areas of wine quality, iris, and heart disease. Subpar performance on the breast cancer dataset, however, made this flaw clear. Both two-class and multiclass datasets were handled with benefits by logistic regression, which also performed well in low- and high-dimensional settings. Algorithm-data-centric, multifaceted ecosystems are compared and ranked. SVM ranked third, exhibiting superior performance in breast cancer, while KNN and Naïve Bayes followed suit, standing out in diverse environments and contexts.

Saravanan and Charan [18] used Convolutional Neural Networks (CNN) and SVM as Machine Learning tools in a dataset for human activity recognition, achieving accuracy rates of 92.46% and 90.19%, respectively. These results indicate the superiority of the CNN classifier for the analysis of human activities. However, it is important to note that Rough Set Theory (RST) is rarely mentioned in reviews of Machine Learning algorithms, despite several studies establishing a solid foundation for its application in Machine Learning problems. Singh and Yao [19] emphasized that previous research addressed pneumonia identification using Machine Learning algorithms, particularly Deep Learning, in a binary classification approach. However, they introduced RST during the machine learning process to mitigate uncertainty, achieving a remarkable accuracy rate of 96.25% in pneumonia detection. Nayak et al. [20] conducted a study using the RST to reveal latent information in imprecise datasets, focusing on the accurate identification of malaria symptoms as conditional attributes.

III. METHODOLOGY

The dataset for this research was taken from scientific studies in healthcare environments and comprises 1,034 patient samples. These samples encompass a broad set of features ranging from toxicity biomarkers, hematological results, and other clinical parameters that are gathered under supervised conditions with careful domain guidance. The dataset was thoroughly cleaned and balanced with sophisticated methods like SMOTE (Synthetic Minority Over-sampling Technique) to make class representations fair and avoid bias while training. The proposed model design is shown in figure 1.

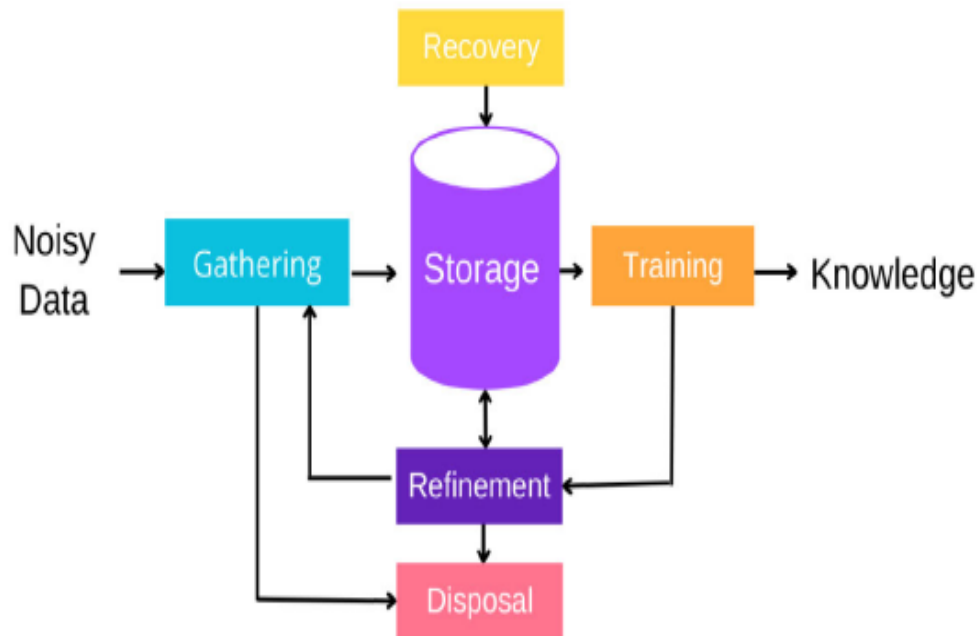


Figure 1. Framework of proposed work

Data research (manual collection) or existing data sources (digital collection) are two ways that data may be acquired. Whichever approach is used, the following pillars need to be taken into account.

Privacy:

Only those in charge of the collection, who should have the necessary technical skills, should have access to data when it is collected manually. Determining access levels is crucial in the context of digital collection in order to guard against confidentiality violations and guarantee adherence to legal requirements for the protection of digital data.

Copyright:

In order to utilise data in compliance with current laws and ethical standards, written consent from the data owner must be obtained.

Quality:

When collecting data by hand, especially when it is not yet available, preparation is essential. To guarantee the quality of data gathering, this may entail standardising questions and answers and preventing repeated answers to the same query. Finding the repositories and confirming that the data are appropriately standardised are crucial in the case of digital collecting, which is predicated on preexisting data.

Using pre-existing information supplied by a domain specialist, digital data was obtained for the PHH project and entered into a spreadsheet. It is important to note that a large percentage of these data points were noise-filled, with just a small number of data points showing regular response patterns.

The data refinement process is a crucial phase in data science that includes crucial steps including resolving missing values, standardisation, and the removal of unnecessary information and change when required. While the RDBMS maintains data privacy and integration, the DRC–SML model suggests using a specially designed form to record the refining process of each recorded data to preserve data preservation and quality over time. Important details including the kind of data (continuous, discretised, multivalued, or null) and its state throughout the refining process—whether it is finished or ought to be discarded—are covered in this form.

Notably, our refining technique is greatly influenced by the type of the data. Discreteization criteria must be established for continuous data to provide defined intervals that are appropriate for further analysis. Documenting the

previously used rules is crucial for discretised data to facilitate process comprehension and replication. Finally, when working with null or multivalued data, it is wise to determine whether it would be feasible to go back to the collection stage. For multivalued data, we recommend breaking the data down into discrete units and giving each unique piece of information a new classification. As a result, the structure for additional analysis is clearer and more useful.

The goal of the suggested machine learning-based model for identifying pesticide poisoning in rural labourers is to employ supervised learning methods to forecast the risk of poisoning based on a number of characteristics. The model starts by gathering extensive data from several sources, such as demographic information (gender, age), environmental parameters (temperature, humidity, pesticide concentration), symptoms (headache, nausea, dizziness), and pesticide exposure history (type, duration, frequency). To guarantee that these data points are appropriate for machine learning, they are then meticulously pre-processed. This entails dealing with missing data, storing categorical variables like pesticide kind and symptoms using techniques like one-hot encoding, and scaling numerical features like age and exposure time to a consistent range. The objective is to produce a clear, organised dataset with properly represented characteristics for model training. Choosing the right machine learning algorithm comes next once the data is prepared. The best algorithm for forecasting pesticide toxicity is determined by evaluating algorithms like Decision Trees, Random Forests, Support Vector Machines (SVM), and Neural Networks. To identify patterns and connections between input characteristics (such as pesticide exposure and symptoms) and the target variable (pesticide poisoning diagnosis), these algorithms are trained on historical data. Metrics including accuracy, precision, recall, and F1-score are used to assess the model after training to make sure it can accurately identify pesticide poisoning in novel instances. Following selection and validation, the optimal model is put into use in the field, allowing medical practitioners to enter pertinent data and get predictions in real time. In addition to aiding in the early diagnosis of pesticide toxicity, our machine learning model lays the groundwork for future advancements through constant data updates and retraining, guaranteeing its accuracy throughout time.

IV. EXPERIMENTAL FINDINGS

A. System Environment

The proposed model is developed using said hardware system configuration and software required to implement the design is shown below.

Hardware System Configuration:

- Processor: Intel i5
- RAM : 8 GB
- Hard Disk : 256 GB
- Key Board : Standard Windows Keyboard
- Mouse : Two or Three Button Mouse
- Monitor : SVGA

Software Requirements:

- Operating system : Windows 7 Ultimate.
- Coding Language: Python.
- Front-End : Python.
- Back-End : Django-ORM
- Designing : Html, css, javascript.
- Data Base : MySQL (WAMP Server).

B. Experimental Results

In the PHH project, we performed CV following these steps:

- We selected samples according to the type of diagnosis, resulting in nine distinct selections.
- Next, we distributed these samples to create four partitions of proportional sizes, each containing approximately the same percentage for each diagnosis type.

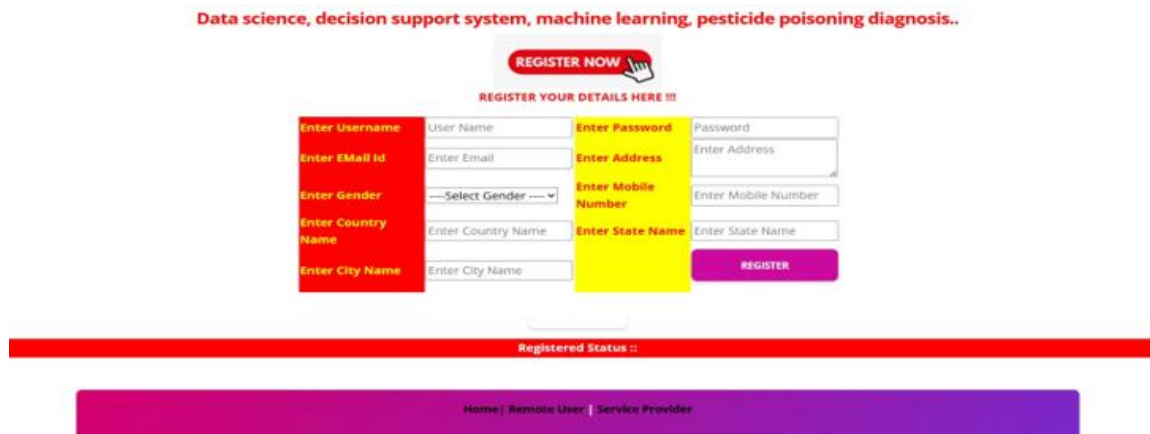
These partitions were used for conducting four distinct trainings, alternating the test partition denominated Tests A, B, C, and D, while the remaining partitions underwent training, as illustrated in following figures.

Data science, decision support system, machine learning, pesticide poisoning diagnosis..



The login interface features a central icon of a person and a padlock. Below it, the text "Login Using Your Account:" is displayed. The form includes two input fields: "User Name" and "Password". A "LOGIN" button is positioned below these fields. At the bottom of the form, there is a link that says "Are You New User !!! REGISTER". A purple navigation bar at the bottom contains the text "Home | Remote User | Service Provider".

Figure 2. Login interface



The registration interface starts with a "REGISTER NOW" button and the text "REGISTER YOUR DETAILS HERE !!!". The form is organized into two columns. The left column has a red header and includes fields for "Enter Username", "Enter EMail Id", "Enter Gender" (with a dropdown menu), "Enter Country Name", and "Enter City Name". The right column has a yellow header and includes fields for "Enter Password", "Enter Address", "Enter Mobile Number", and "Enter State Name". A "REGISTER" button is located at the bottom right of the form. Below the form is a red bar with the text "Registered Status ::". A purple navigation bar at the bottom contains the text "Home | Remote User | Service Provider".

Figure 3. User registration interface



The "ENTERING THE DETAILS" interface is set against a background image of a green field. It features a red sidebar on the left with labels for "Enter Fid", "Enter Age", "Enter Gender" (with a dropdown menu), "Enter PesticideName", "Enter Crop", "Enter State", "Enter Location", "Enter Area", and "Enter AreaUnits". Each label is followed by an input field. Below the input fields is a "Predict" button. At the bottom, there is a red box labeled "Detected Pesticide Poisoning Diagnosis Status".

Figure 4. Entering the details

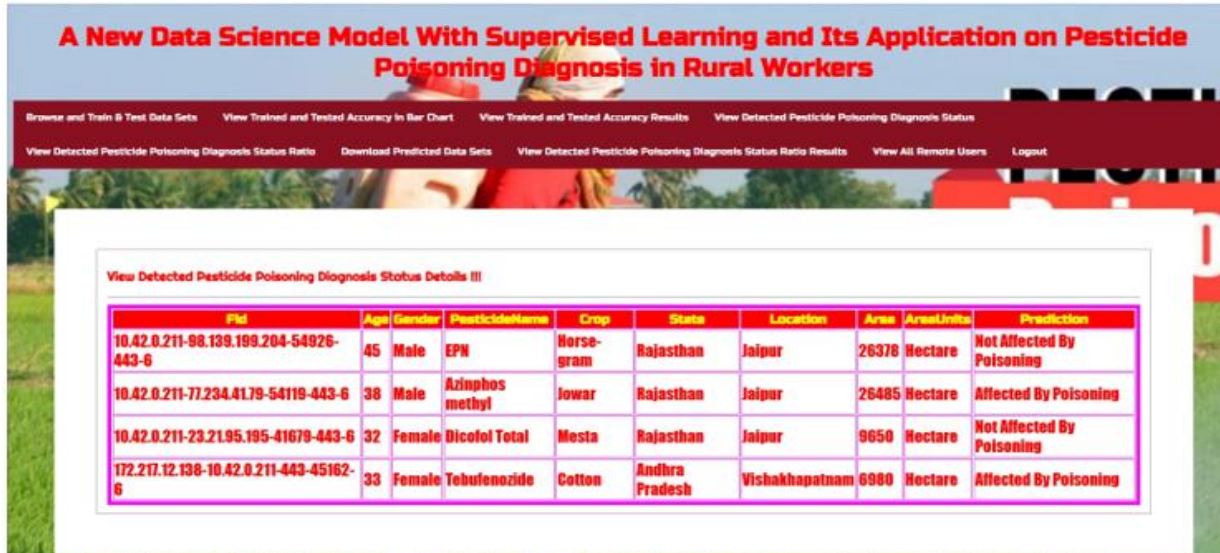


Figure 5. Interface after entering details



Figure 6. Output Prediction details

V. CONCLUSION

The DRC–SML model demonstrated its effectiveness in addressing all stages of the KDD process according to the hierarchy of Data Science. The model learned with XGBoost, specifically, performed extremely well with 99.61% accuracy using only 27 decision rules. This minimalistic approach to decision-making makes the model extremely interpretable and implementable in actual healthcare settings, particularly where computational power is constrained. The project results demonstrate how combining domain knowledge with strong machine learning methods can produce effective healthcare solutions. The DRC–SML approach not only yielded extremely accurate diagnostic outcomes but also presented a scalable and flexible model for future Data Science implementation in other healthcare and public safety applications.

In future work, it is recommended: To explore the connection of the DRC–SML model with object-oriented databases and graph-oriented databases, allowing for greater flexibility in data storage and retrieval. Expanding the scope by incorporating the necessary documentation into the DRC–SML modeling for tracking additional machine

learning methods. This approach will enable data scientists to assess and choose the most suitable method based on the specific characteristics of their problems, allowing for comparative analyses on the same dataset. To improve the accessibility and practical utility of the model, a possible extension would be to create an API to implement a mobile data collection application.

REFERENCES

- [1] C. J. Cremin, S. Dash, and X. Huang, “Big data: Historic advances and emerging trends in biomedical research,” *Current Res. Biotechnol.*, vol. 4, pp. 138–151, Jan. 2022.
- [2] J. Saltz and I. Krasteva, “Current approaches for executing big data science projects a systematic literature review,” *PeerJ Comput. Sci.*, vol. 8, p. e862, Feb. 2022.
- [3] J. D. Kelleher and B. Tierney, *Data Science*. Cambridge, MA, United States: MIT Press, 2018.
- [4] C. Silva, M. Saraee, and M. Saraee, “Data science in public mental health: A new analytic framework,” in *Proc. IEEE Symp. Comput. Commun. (ISCC)*, Jun. 2019, pp. 1123–1128.
- [5] S. Jain, “Comprehensive survey on data science, lifecycle, tools and its research issues,” in *Proc. Int. Conf. Mach. Learn., Big Data, Cloud Parallel Comput.*, vol. 1, May 2022, pp. 838–842.
- [6] T. D. Bie, L. D. Raedt, J. Hernández-Orallo, H. H. Hoos, P. Smyth, and C. K. I. Williams, “Automating data science: Prospects and challenges,” *Commun. ACM*, vol. 65, no. 2, pp. 76–87, 2022.
- [7] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, pp. 281–296, Dec. 2019.
- [8] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning About Data*, 1st ed. Dordrecht, The Netherlands: Springer, 1991.
- [9] D. P. Acharjya and A. Abraham, “Rough computing—A review of abstraction, hybridization and extent of applications,” *Eng. Appl. Artif. Intell.*, vol. 96, Nov. 2020, Art. no. 103924.
- [10] I. H. Sarker, “Data science and analytics: An overview from data-driven smart computing, decision-making and applications perspective,” *Social Netw. Comput. Sci.*, vol. 2, no. 5, p. 377, Sep. 2021.
- [11] A. Peña-Fernández, M. Peña, M. Lobo, and M. Evans, “Interventions to enhance the teaching of toxicology at a U.K. University,” in *Proc. EDULEARN Conf.*, Palma, Spain, Jul. 2018, pp. 7126–7130.
- [12] B. Mondal, *Artificial Intelligence: State of the Art*. Cham, Switzerland: Springer, 2020, pp. 389–425, doi: 10.1007/978-3-030-32644-9.
- [13] J. Saltz, N. Hotz, D. Wild, and K. Stirling, “Exploring project management methodologies used within data science teams,” in *Proc. Americas Conf. Inf. Syst.*, 2018, pp. 1–5.
- [14] A. K. Laturiw and Y. A. Singgalen, “Sentiment analysis of raja ampat tourism destination using CRISP-DM: SVM, NBC, DT, and k-NN algorithm,” *J. Inf. Syst. Informat.*, vol. 5, no. 2, pp. 518–535, May 2023.
- [15] T. Kaewwiset, P. Temdee, and T. Yooyativong, “Employee classification for personalized professional training using machine learning techniques and SMOTE,” in *Proc. Joint Int. Conf. Digit. Arts, Media Technol. ECTI Northern Sect. Conf. Electr., Electron., Comput. Telecommun. Eng.*, Mar. 2021, pp. 376–379.
- [16] B. Abdualgalil and S. Abraham, “Applications of machine learning algorithms and performance comparison: A review,” in *Proc. Int. Conf. Emerg. Trends Inf. Technol. Eng.*, Feb. 2020, pp. 1–6.
- [17] V. Sarveshwaran, J. J. Jayakanth, and Y. Renu, “Comparative analysis of diverse classification algorithms of machine learning by using various quality metrics,” in *Proc. IEEE 5th Int. Conf. Cybern., Cognition Mach. Learn. Appl. (ICCCMLA)*, Oct. 2023, pp. 551–556.
- [18] M. S. Saravanan and S. Charan, “Prediction of insufficient accuracy for human activity recognition using convolutional neural network in compared with support vector machine,” in *Proc. 5th Int. Conf. Contemp. Comput. Informat. (IC3I)*, Dec. 2022, pp. 1915–1919.
- [19] Mudunuri, L. N. R., Hullurappa, M., Vemula, V. R., & Selvakumar, P. (2025). AI-powered leadership: Shaping the future of management. In *Navigating Organizational Behavior in the Digital Age With AI* (pp. 127-152). IGI Global Scientific Publishing.
- [19] S. Singh and J. Yao, “Pneumonia detection with game-theoretic rough sets,” in *Proc. 20th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2021, pp. 1029–1034.
- [20] S. K. Nayak, S. K. Pradhan, S. Mishra, S. Pradhan, and P. K. Pattnaik, “Rough set technique to predict symptoms for malaria,” in *Proc. 8th Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, Mar. 2021, pp. 312–317.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  ijirset@gmail.com



www.ijirset.com

Scan to save the contact details