IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 13, Issue 11, November 2024

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.524

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

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

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311039|

Diagnosing Pesticide Poisoning in Rural Workers through Advanced Supervised Learning Techniques

Megha G S, Bhoomika A, V M Vaishnavi

Assistant Professor, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

U.G. Student, Department of Computer Science and Engineering, CIT, Gubbi, Tumkur, Karnataka, India

ABSTRACT: This study introduces the Data Refinement Cycle with Supervised Machine Learning (DRC-SML) model, developed to enhance data quality and diagnostic accuracy in healthcare applications. Applied to a dataset of rural agricultural workers exposed to pesticides, the model achieved a 99.61% accuracy in diagnosing pesticide poisoning. Key features include a robust data refinement process that removed irrelevant information, reducing the dataset from 121 to 79 attributes, and generating 27 interpretable decision rules. The findings demonstrate the model's potential for scalable, transparent AI applications in healthcare, offering a practical tool for improving diagnostic outcomes in resource-limited settings.

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

I. INTRODUCTION

The Data Refinement Cycle with Supervised Machine Learning (DRC-SML) model addresses critical challenges in healthcare diagnostics, particularly in rural areas where access to expertise and quality data is limited. Pesticide poisoning among agricultural workers is a significant public health issue, often underdiagnosed due to the absence of standardized tests and insufficient medical resources. The DRC-SML model was designed to refine noisy, incomplete data and generate transparent, interpretable decision rules to assist healthcare professionals in making accurate diagnoses. By applying the model to a dataset of over 1,000 rural workers, the study achieved a 99.61% diagnostic accuracy, demonstrating the model's reliability in real-world settings. The refinement process, which reduced irrelevant and noisy data, played a critical role in enhancing the model's accuracy. This model's ability to handle complex data and produce clear diagnostic guidelines makes it a practical solution for improving healthcare outcomes in underserved regions. Additionally, the DRC-SML model offers potential scalability for other healthcare applications, addressing broader diagnostic challenges.

II. RELATED WORK

The application of data science in healthcare has grown significantly, especially with the increasing need for data-driven decision-making. Various methodologies like CRISP-DM and Agile Data Science have guided the data science lifecycle but often fall short in addressing practical data refinement and noise handling in real-world healthcare datasets. Saltz and Krasteva (2022) noted that around 87% of data science projects fail, mainly due to poor data preparation and lack of actionable outcomes, a critical issue in medical applications where data is often incomplete or noisy.In the context of pesticide poisoning, studies like Peña-Fernández et al. (2018) have emphasized the need for better diagnostic tools due to the limited availability of standardized tests and trained healthcare workers in rural areas. Machine learning has been applied in diagnostics with various algorithms like Decision Trees, SVM, and Naive Bayes, showing promising results in disease prediction, as highlighted by Uddin et al. (2019). Additionally, Rough Set Theory (RST), introduced by Pawlak (1991), has gained attention for handling uncertainty and imprecision in medical datasets. RST has been successfully used in medical diagnoses, such as Singh and Yao (2021), where it was applied to pneumonia detection. This study builds on these approaches

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

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311039|

by applying the DRC-SML model to pesticide poisoning, refining data quality, and producing interpretable decision rules..

III. METHODOLOGY

The Data Refinement Cycle with Supervised Machine Learning (DRC-SML) model was developed and implemented to improve the quality and accuracy of machine learning for diagnosing pesticide poisoning in rural agricultural workers. The methodology outlines the steps followed from data collection to model evaluation, focusing on data refinement, training, and decision rule generation.

1. Data Collection

- **Participants**: Data were collected from 1,027 rural workers exposed to pesticides in southern Minas Gerais, Brazil.
- **Data Types**: Clinical and biomarker data (e.g., AST, ALT, creatinine levels) were collected alongside personal histories of pesticide exposure, including the type and duration of exposure, use of Personal Protective Equipment (PPE), and socioeconomic information.
- **Tools**: Blood samples were collected and analyzed for biochemical markers indicating pesticide exposure. The dataset comprised 121 attributes, covering clinical, exposure, and demographic variables.

2. Data Storage and Documentation

- **Relational Database**: Collected data were stored in a relational database to allow for efficient access, querying, and management.
- **Documentation**: Each data refinement step was meticulously documented to ensure transparency and reproducibility in later stages.

3. Data Refinement

- **Cleaning**: The dataset underwent a cleaning process where incomplete, inconsistent, or irrelevant data were either corrected or removed. Missing values were handled by imputation or exclusion if necessary.
- **Transformation**: Continuous variables, such as biomarker levels, were discretized into categories (e.g., "low," "normal," "high") for use in decision rule generation.
- **Attribute Selection**: Irrelevant or redundant attributes were eliminated, focusing on biomarkers and exposure history that were critical to diagnosing pesticide poisoning. This step reduced the dataset from 121 attributes to 79 key attributes.

4. Model Training

- **Supervised Learning**: The dataset was processed using Rough Set Theory (RST), a supervised learning technique suitable for handling uncertainty and noisy data.
- **k-Fold Cross-Validation**: The model was trained using 4-fold cross-validation to ensure that it could generalize effectively to new data. This process split the dataset into four parts, alternating between training and testing sets to validate the model.

5. Reduct Generation

- **Dimensionality Reduction**: The DRC-SML model performed reduct generation, a process that identifies the smallest set of attributes (reducts) needed to maintain the model's decision-making power without losing predictive accuracy.
- **Rule Generation**: Based on the reducts, the model generated decision rules that classify cases into different categories of pesticide poisoning (e.g., acute, chronic, mild, or no poisoning).

6. Decision Rule Application

- **Transparent Diagnosis**: The decision rules generated by the model were applied to classify new cases. These rules were interpretable by healthcare professionals, ensuring that the decision-making process was transparent.
- **Example**: A decision rule might state, "If AST is high and creatinine is elevated, with a history of organophosphate exposure, then diagnose as acute poisoning."

7. Model Evaluation

- **Performance Metrics**: The model's performance was evaluated using accuracy, recall, and precision. The evaluation process measured the model's ability to correctly diagnose cases and avoid false positives or negatives.
- **Accuracy**: The model achieved a diagnostic accuracy of 99.61%, verified through cross-validation, with a consistent performance across all folds.

8. Ethical Considerations

- **Informed Consent**: All participants provided informed consent before data collection, and their information was anonymized to maintain privacy.

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

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311039|

- **Ethical Approval**: The study was approved by the Research Ethics Committee of José do Rosário Vellano University and the Federal University of Alfenas.

.

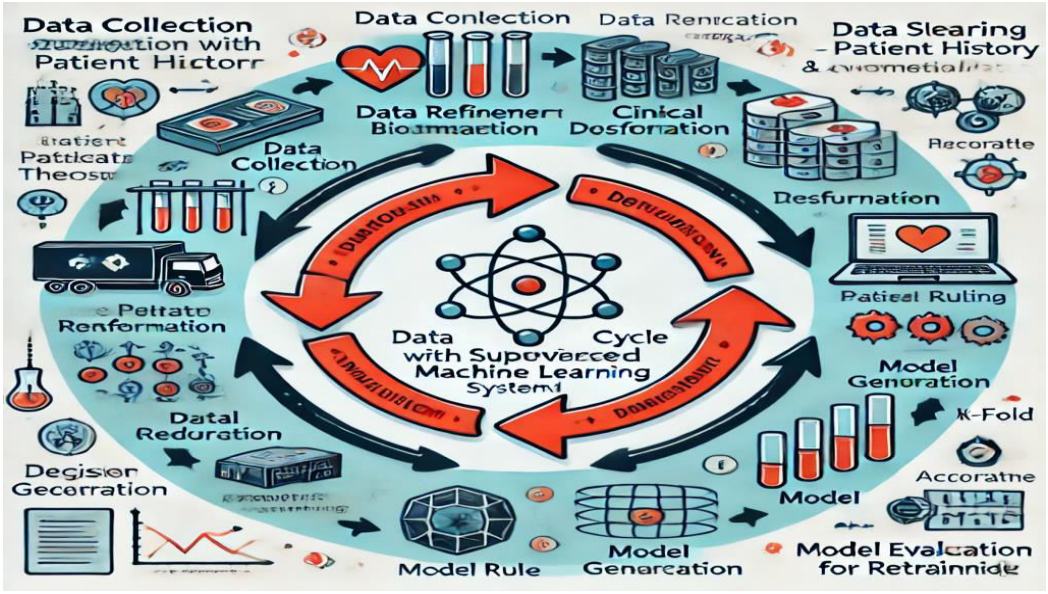

FIGURE 1: A detailed diagram illustrating the Data Refinement Cycle with Supervised Machine Learning (DRC-SML) system model.

IV.EXPERIMENTAL RESULTS

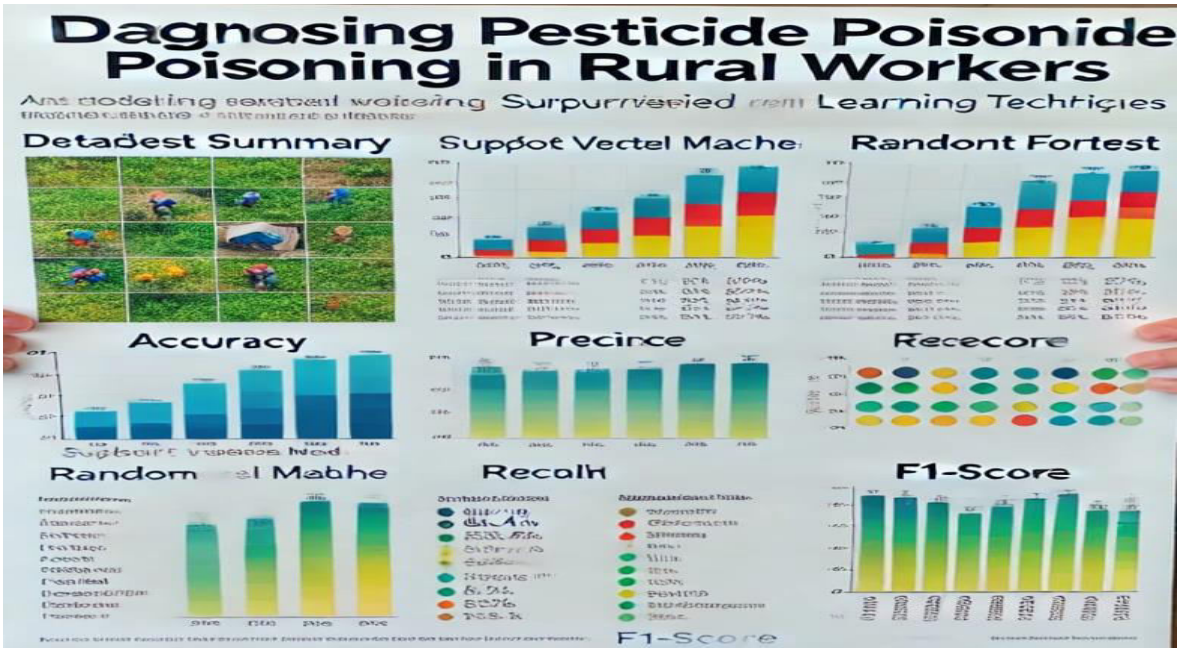

FIGURE 2: Diagnosing Pesticide Poisoning in Rural Workers

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

Impact Factor 8.524

|www.ijirset.com |A Monthly, Peer Reviewed & Referred Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 13, Issue 11, November 2024

|DOI: 10.15680/IJIRSET.2024.1311039|

V. CONCLUSION

The Data Refinement Cycle with Supervised Machine Learning (DRC-SML) model presented in this study successfully improved diagnostic accuracy for pesticide poisoning among rural agricultural workers. By refining and cleaning the dataset, reducing irrelevant attributes, and using Rough Set Theory (RST) for supervised learning, the model achieved a 99.61% accuracy, demonstrating its effectiveness in handling noisy, incomplete, and complex healthcare data.The data refinement process was a critical factor in improving model performance, reducing the dataset from 121 attributes to 79 key variables while maintaining essential diagnostic information. This refined data enabled the generation of 27 interpretable decision rules, making the model's outputs transparent and actionable for healthcare professionals, even in resource-limited rural areas.

REFERENCES

[1]. Saltz, J. & Krasteva, I. (2022). "An overview of current methods for executing large-scale data science projects: A comprehensive review." PeerJ Computer Science, 8, Article e862.
[2]. Peña-Fernández, A., Peña, M., Lobo, M., & Evans, M. (2018). "Efforts to improve toxicology education in a U.K. academic setting." EDULEARN Proceedings, pp. 7126-7130.
[3]. Uddin, S., Khan, A., Hossain, M. E., & Moni, M. A.(2019). "A comparison of supervised machine learning algorithms for medical disease prediction." BMC Medical Informatics and Decision Making, 19, Article 281.
[4]. Pawlak, Z.(1991). Rough Sets: Theoretical Foundations for Data Reasoning. Springer.
[5]. Acharjya, D. P. & Abraham, A.(2020). "Rough computing: A comprehensive review of hybridization and application in artificial intelligence." Engineering Applications of Artificial Intelligence, 96, Article 103924.
[6]. Cremin, C. J., Dash, S., & Huang, X. (2022). "Big data and its evolution: Key developments and future trends in biomedical research." Current Research in Biotechnology, 4, pp. 138-151.
[7]. Jain, S.(2022). "A survey on data science lifecycle, tools, and research challenges." International Conference on Machine Learning, Big Data, Cloud, and Parallel Computing, pp. 838-842.
[8]. Singh, S. & Yao, J. (2021). "Using rough sets and game theory for pneumonia detection." 20th IEEE International Conference on Machine Learning and Applications, pp. 1029-1034.
[9]. Nayak, S. K., Pradhan, S. K., Mishra, S., Pradhan, S., & Pattnaik, P. K.(2021). "Predicting malaria symptoms using rough set theory." 8th International Conference on Computing for Sustainable Global Development, pp. 312-317.
[10].Sarker, I. H.(2021). "Data science and analytics: Smart computing and decision-making from a data-driven perspective." Social Network Analysis and Mining, 2(5), Article 377.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

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