Proceedings of the 7th International Conference on Inventive Computation Technologies (ICICT 2024)
Xplore Part Number : CFP24F70-ART ; ISBN : 979-8-3503-5929-9

2024 International Conference on Inventive Computation Technologies (ICICT) | 979-8-3503-5929-9/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICICT60155.2024.10544901

Comprehensive Analysis of Multiple Diseases using Machine Learning

[1] Dr. D. Rosy Salomi Victoria,
Professor,
Department of Computer Science and Engineering,
Jaya Engineering College,
Thiruninravur, Tamil Nadu
drosysalomi@gmail.com

[2]T. Subhishwaran
Department of CSE
St. Joseph's College of Engineering,
Chennai, Tamil Nadu.
subhish2208gmail.com

[3]P. Velan
Department of CSE
St. Joseph's College of Engineering,
Chennai, Tamil Nadu.
velanvela5@gmail.com

***Abstract*—In the contemporary landscape of the healthcare industry, the pervasive integration of computer-based technology has resulted in the prolific accumulation of electronic health data. However, this abundance of information poses a significant challenge for medical professionals, impeding their ability to accurately analyze symptoms and promptly detect diseases. Addressing this issue, supervised Machine Learning (ML) algorithms have emerged as a promising solution, showcasing superior performance compared to traditional diagnostic methods. This literature review aims to discern patterns in disease diagnosis across various supervised ML models, with a specific focus on predicting fungal infections, diabetes, jaundice, malaria, and heart diseases. The model faces challenges in managing complex disease data, addressing class imbalances, and optimizing feature selection. Disease datasets present complexities in symptom-disease relationships, while imbalances in disease occurrences impact model accuracy. Among the myriad of algorithms explored, particular attention is directed towards the efficacy of Random Forest and Decision Trees. The proposed model aims to develop robust techniques for handling complex disease data, addressing class imbalances, and optimizing feature selection. These objectives target the improvement of disease prediction models, enhancing their accuracy and reliability for practical applications. Through an in-depth analysis of performance indicators, this study intends to contribute valuable insights into enhancing the diagnostic capabilities of healthcare professionals, facilitating early identification of high-risk conditions, and ultimately improving patient outcomes.**

***Keywords*–Machine Learning; Random Forest; Decision Tree; Multi-Disease Diagnosis; Proactive Diseases.**

I. INTRODUCTION

The landscape of modern healthcare, marked by the widespread adoption of computer-based technology, has ushered in an era of unprecedented electronic data accumulation. This surge in data holds immense potential for improving disease diagnosis and management. However, medical professionals grapple with the formidable challenge of effectively analyzing vast datasets, hindering early detection and accurate interpretation of symptoms.

In response to this challenge, supervised machine learning (ML) algorithms have emerged as beacons of promise, demonstrating their superiority over conventional methods in diagnosing a spectrum of diseases. This literature review aims to unravel discernible patterns in disease diagnosis, focusing on five prevalent health conditions: fungal infection, diabetes, jaundice, malaria, and heart disease. These diverse ailments pose unique challenges, demanding advanced diagnostic tools and methodologies. Among the array of supervised ML algorithms, particular attention is directed towards random forests and decision trees, both recognized for their efficacy in medical diagnostics. The model's scope encompasses the prediction and early identification of these diseases through the implementation of supervised ML algorithms. Notably, the study incorporates data from various sources, including patient records, clinical trials, and genetic testing, to enable a comprehensive analysis. The experimental results show that the proposed method outperformed conventional machine learning algorithms by achieving an accuracy of 1.0.

II. RELATED WORKS

A concise overview of diabetes prediction [1] focuses on integrating machine learning and ontology. This survey examines the existing research on predictive modeling and ontology-driven approaches, highlighting their importance for enhancing prediction and accuracy in diabetes management. The insights into heart disease prediction [2] offer effective feature engineering with machine learning. This study focuses on optimizing feature selection techniques to enhance prediction accuracy, contributing to advancements in cardiovascular health assessment using data-driven approaches.

The classification algorithms' performance on maintainable applications [3] provides benchmarks for Advances in Information Communication and Cyber security.

This study contributes to understanding algorithm efficacy in handling maintainability tasks, crucial for enhancing software development practices and cyber security measures. A method to enhance human heart disease prediction [4] in Mobile Information Systems is proposed. This approach aims to improve prediction accuracy, potentially aiding in early detection and preventive healthcare strategies for cardiovascular diseases.

A comprehensive survey on secure and robust machine learning techniques [5] is conducted in healthcare. This study explores various methodologies and advancements aimed at ensuring data privacy, integrity, and model reliability in healthcare applications, addressing critical concerns in adopting machine learning for medical purposes. A study comparing supervised machine learning algorithms for disease prediction [6] aims to assess their performance. This outlines the experimental setup, encompassing algorithm selection, preprocessing methods, feature selection, and evaluation metrics. Various classification algorithms on three disease-related databases [7] from the University of California, Irvine repository employ backward modeling for feature selection. This study underscores the effectiveness of machine learning techniques in early disease detection. The rising incidence of heart strokes [8] in young individuals is addressed by proposing a mobile application for early detection based on fundamental symptoms. Utilizing neural networks, known for accuracy, the system aims to predict heart disease risk conveniently without extensive testing and empower individuals with readily available information.

The detection of associations between diseases and genes [9] through genome-wide association studies presents challenges due to vast potential mutations and high false-positive rates. Computational approaches, particularly network-based methodologies, offer cost-effective and complementary evidence for disease-gene associations, capturing intricate molecular interactions. The dermatological diseases [10] pose challenges due to their complexity and subjective diagnosis methods. These proposed automated methods use diverse lesion images for early disease diagnosis, employing convolutional neural networks and support vector machines (SVM) for robustness.

The disease module hypothesis [11] suggests that cellular components associated with the disease tend to cluster together within the human interaction, but the incomplete nature of this map and limited understanding of disease-associated genes raise questions about its applicability. Network medicine [12] offers a systematic approach to understanding disease mechanisms, identifying modules and pathways, and uncovering connections between diverse pathological phenotypes.

Despite the advancements in identifying disease-associated genes, understanding the intricate cellular mechanisms [13] remains limited, especially for complex traits and Mendelian diseases. The prevailing perspective suggests diseases stem from disruptions in molecular networks, where genes linked to similar diseases cluster together. A network integrating disorders and disease-associated genes [14] is connected through established disorder-gene associations, offering a comprehensive view within a unified graph-theoretic framework. This structure enables the exploration of phenotype-disease gene associations, revealing shared genetic bases across diseases. Deep learning's applications in biomedical tasks like patient classification [15], understanding biological processes and treatment, and assessing its transformative potential and unique challenges in this domain are utilized.

The benefits highlighted from the survey encompass improved disease management through enhanced prediction accuracy via machine learning and ontology integration, refined feature engineering techniques, and secure machine learning practices in healthcare. Furthermore, supervised learning algorithms prove effective in detecting diseases early, while computational methods offer cost-efficient approaches to identifying disease-gene associations and facilitating early diagnosis. Nevertheless, persisting challenges such as incomplete disease maps and a limited understanding of disease mechanisms underscore the need for ongoing research. Additionally, the adoption of deep learning in biomedical tasks presents significant opportunities alongside unique challenges, emphasizing the importance of careful consideration in practical implementations.

III. METHODOLOGY

A diverse dataset encompassing relevant health parameters, demographic information, and historical records of individuals diagnosed with diseases is meticulously collected. The emphasis is placed on sourcing data from reliable healthcare repositories and clinical studies. Thorough data pre-processing methods are utilized to manage absent data, standardize characteristics, and deal with any anomalies. This involves data cleaning, feature scaling, and encoding categorical variables to ensure the dataset is conducive to effective model training. The dataset is partitioned into training and testing sets, with 80% for training and 20%, to testing to facilitate robust model evaluation. Stratified sampling ensures a balanced representation of positive and negative cases.

Cross-validation techniques are applied during the training phase to enhance the model’s generalizability.

Supervised machine learning algorithms, including K Nearest Neighbor (KNN) and Decision Trees, are implemented. These algorithms are chosen for their proven efficacy in multiple disease prediction. By analyzing similarities between data points, KNN effectively classifies diseases for accurate prediction. It complements other algorithms in the system, contributing to enhanced disease prediction performance. The model's predictive performance is evaluated using metrics such as accuracy, recall, precision, and F1-score.

Figure 1 illustrates the system architecture, utilizing machine learning to preprocess a symptoms dataset, followed by model training on an 80% training set and a 20% test set using a Django server.

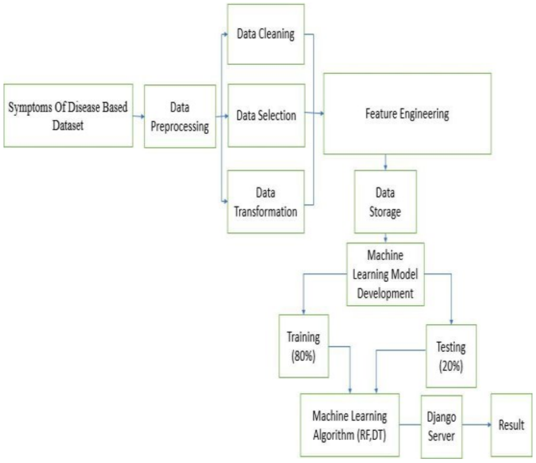


Figure 1 System Architecture Diagram

The disease prediction process involved sourcing symptoms data from a Kaggle repository, followed by thorough data preprocessing and cleaning. Feature engineering and selection optimized the dataset for machine learning. Multiple algorithms, including KNN, Random Forest, and SVM, were trained and tested, culminating in a Django server for user-friendly access to accurate disease predictions.

IV. MATERIALS AND METHODS

A. DATASET COLLECTION

Data collection is a systematic process involving the gathering, compilation, and organization of information relevant to a particular study. The process adheres to ethical considerations, emphasizing data quality, diversity, and representatives to establish a solid foundation for subsequent analyses and machine learning model development. Creating a dataset for multiple disease prediction involves several steps, and it's important to ensure that the data is collected and used ethically. Leveraging a carefully collected dataset from Kaggle, this study seeks to advance the field of healthcare diagnostics by implementing state-of-the-art machine learning techniques.

- Fungal infection: itching, skin rash, dyschromic patches, etc.
- Diabetes: blurred and distorted vision, increased appetite, polyuria, etc.
- Jaundice: yellowish skin, abdominal pain, etc.

Figure depicts a heap map utilized in machine learning for disease prediction, enhanced pattern recognition, and correlation analysis in the dataset.

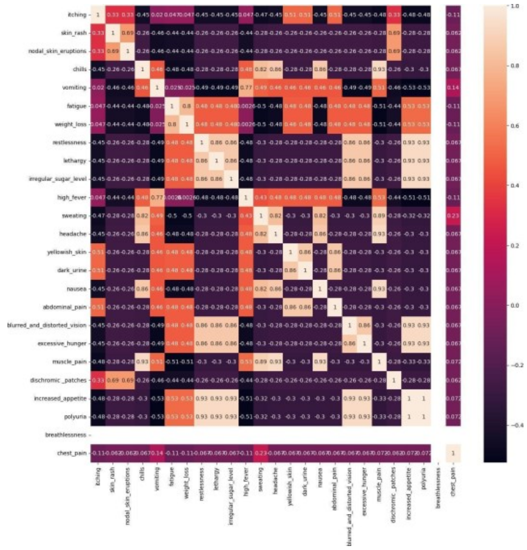


Figure 2 Heat map

B. DATA PREPROCESSING

Data preprocessing is a critical phase where the raw dataset undergoes systematic transformations to enhance its suitability for machine learning models.

In the preprocessing phase of the study, the dataset containing binary-encoded symptoms for various diseases was initially examined for missing values, which were then addressed using imputation techniques such as mode substitution. Redundant features, if present, were identified and eliminated to streamline the dataset and improve computational efficiency. To address class imbalances among diseases, oversampling, undersampling, or synthetic data generation methods were applied as needed.

Data reduction involves various techniques to streamline datasets. This includes selecting only essential features, reducing dimensionality while preserving crucial information, and employing sampling methods to maintain representativeness while reducing size. Additionally, techniques like data aggregation, binning, and feature extraction simplify data representation.

Figure illustrates the data preprocessing steps, encompassing data cleaning, integration, transformation, reduction, and discretization.

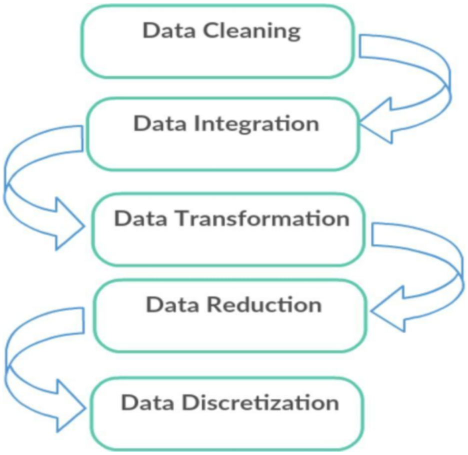


Figure 3 Data Preprocessing

C. DATA SEPARATION

Data separation is a pivotal preprocessing step where the dataset is methodically divided into training and testing subsets. In this disease prediction model, an 80-20 split has been employed, allocating 80% of the data for training the machine learning models and reserving the remaining 20% for testing the model's predictive capabilities. This partitioning ensures the model's exposure to a diverse range of examples during training while evaluating its generalization on unseen data during testing. The 80-20 split is chosen to strike a balance between model learning and assessment precision.

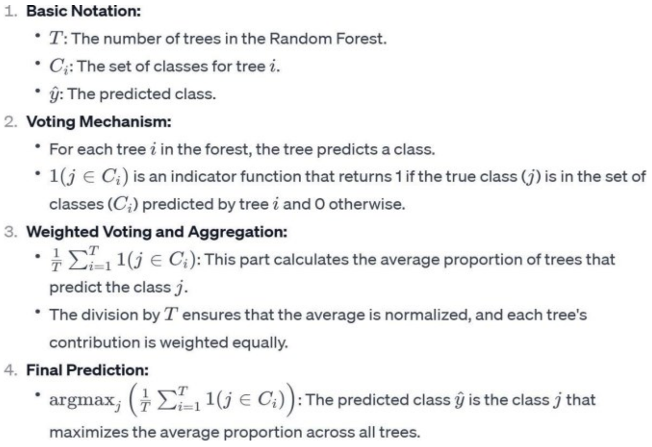
D. MULTIPLE DISEASE DIAGNOSES BY RANDOM FOREST

The Random Forest algorithm's robustness in handling complex relationships between symptoms and diseases, coupled with its ability to effectively manage feature importance and handle high-dimensional data, made it an ideal candidate. Moreover, Random Forest's ensemble approach and inherent ability to mitigate overfitting made it well-suited for the task at hand. Through rigorous evaluation and comparison with other algorithms, Random Forest emerged as the optimal choice, offering a balance of accuracy, scalability, and interpretability for the disease prediction task.

The Random Forest Equation (1) plays a pivotal role

$$\hat{y} = \operatorname{argmax}_j \left(\frac{1}{T} \sum_{i=1}^{T} 1(j \in C_i) \right) \tag{1}$$

in a voting mechanism for predicting the outcome.



1. **Basic Notation:**
   - T: The number of trees in the Random Forest.
   - C_i: The set of classes for tree i.
   - $\hat{y}$: The predicted class.
2. **Voting Mechanism:**
   - For each tree i in the forest, the tree predicts a class.
   - $1(j \in C_i)$ is an indicator function that returns 1 if the true class (j) is in the set of classes (C_i) predicted by tree i and 0 otherwise.
3. **Weighted Voting and Aggregation:**
   - $\frac{1}{T} \sum_{i=1}^{T} 1(j \in C_i)$: This part calculates the average proportion of trees that predict the class j.
   - The division by T ensures that the average is normalized, and each tree's contribution is weighted equally.
4. **Final Prediction:**
   - $\operatorname{argmax}_j \left(\frac{1}{T} \sum_{i=1}^{T} 1(j \in C_i) \right)$: The predicted class $\hat{y}$ is the class j that maximizes the average proportion across all trees.

Figure 4 depicts the Random Forest algorithm, training the dataset to construct decision trees and make predictions for enhanced model performance.

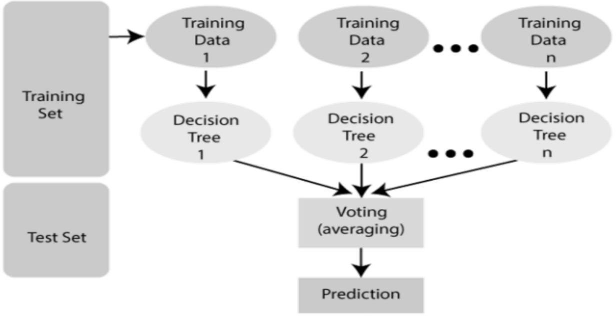


Figure 2 Random Forest Algorithm

The algorithm's ability to handle complex relationships and deliver accurate predictions makes it a suitable choice.

E. WEB DEVELOPMENT

The backend architecture is constructed using Django, a powerful Python web framework. This integration seamlessly incorporates machine learning models, enabling efficient handling of data processing, requests, and model execution. Django's versatility ensures a dynamic and responsive backend that effectively supports disease prediction functionalities. HTML provides the foundational structure of the web application, while CSS enhances the overall presentation and user interface aesthetics. This combination creates an intuitive and visually pleasing front end, ensuring a seamless and engaging user experience for individuals interacting with the disease prediction system.

V. RESULTS & DISCUSSION

The multiple disease prediction model, encompassing the prediction of heart disease, diabetes, jaundice, malaria, and fungal diseases, yielded promising outcomes. Employing machine learning algorithms, particularly Random Forest, the model achieved noteworthy accuracy levels.

Figure 5 illustrates a confusion matrix, vital for assessing a classification algorithm by summarizing key metrics like accuracy, precision, and recall.

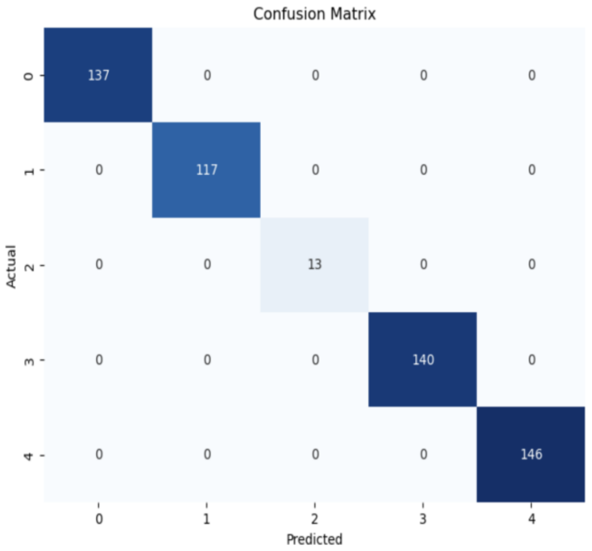


Figure 3 Confusion Matrix

The accuracy of Kernel SVM, K-Nearest Neighbors, Decision Tree, Random Forest, and Naive Bayes models in a concise visual representation is shown in

Figure 4.

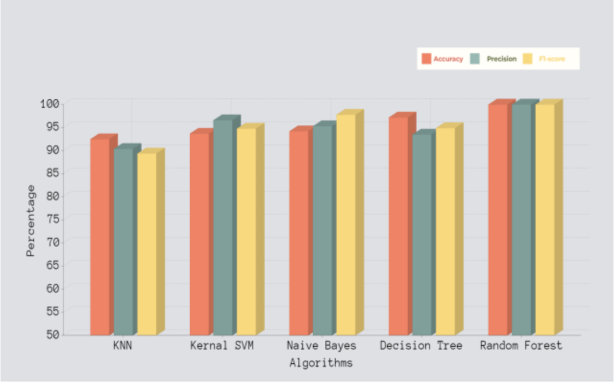


Figure 4 Performance of Models

The Random Forest algorithm demonstrated exceptional performance, achieving perfect scores of 1.0 across all metrics, including accuracy, precision, and F1-score, even after cross-validation. These results underscore the algorithm's robustness and reliability. While highly promising, further exploration of optimization strategies can enhance its already impressive predictive power.

Table 1 illustrates the evaluation metrics, including accuracy, precision, and F1 score, providing a comprehensive assessment of the model's performance.

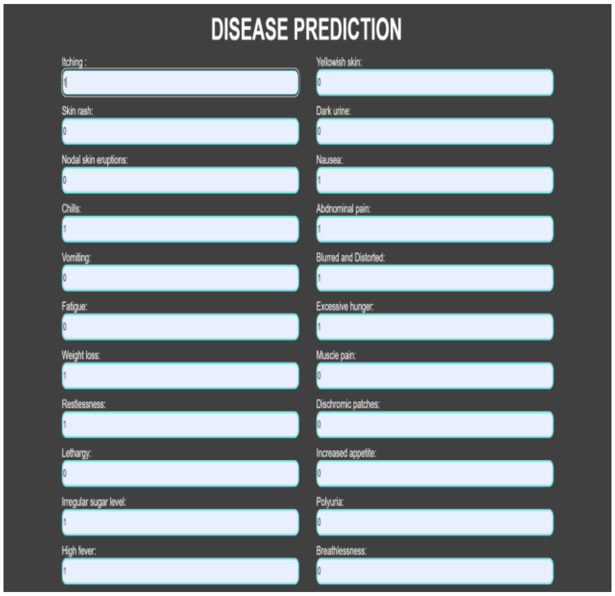
Table 1

Results Comparison

Algorithms	Accuracy	Precision	F1-score
Random forest	1.000	1.000	1.000
Decision tree	0.972	0.935	0.949
Naive Bayes	0.942	0.953	0.978
Kernel SVM	0.937	0.966	0.948
KNN	0.925	0.904	0.894

The performance comparison of various machine learning models reveals that Random Forest achieved perfect scores of 1.000 across all metrics, showcasing its superior predictive accuracy and precision. Decision Tree follows closely behind with high scores in accuracy (0.972) and precision (0.935), although slightly lower than Random Forest. Naive Bayes also demonstrates strong performance, with high scores in precision (0.953) and F1-score (0.978), indicating its effectiveness in classification tasks. Kernel SVM and KNN exhibit slightly lower performance compared to the other models but still maintain respectable scores, highlighting their relevance in diverse predictive scenarios.

Figure 5 depicts the user interface of the model with various symptoms, using which the various possible diseases will be predicted.



DISEASE PREDICTION

Itching : 1
Skin rash: 0
Nodal skin eruptions: 0
Chills: 1
Vomiting: 0
Fatigue: 0
Weight loss: 1
Restlessness: 1
Lethargy: 0
Irregular sugar level: 1
High fever: 1

Yellowish skin: 0
Dark urine: 0
Nausea: 1
Abdnominal pain: 1
Blurred and Distorted: 1
Excessive hunger: 1
Muscle pain: 0
Dischromic patches: 0
Increased appetite: 0
Polyuria: 0
Breathlessness: 0

Figure 5 User Interface

When the malaria symptoms such as chills, vomiting, high fever, sweating, headache, nausea and muscle pain have a value of 1, Figure 6 will be displayed.

YOU ARE AFFECTED BY MALARIA

Figure 6 Malaria

When the diabetes symptoms such as fatigue, weight loss, restlessness, lethargy, irregular sugar level, blurred and distorted vision, excessive hunger, increased appetite, and polyuria have a value of 1, Figure 7 will be displayed.

YOU ARE AFFECTED BY DIABETES

Figure 7 Diabetes

When the jaundice symptoms such as itching, vomiting, fatigue, weight loss, high fever, yellowish skin, dark urine, and abdominal pain have a value of 1, Figure 8 will be displayed.

YOU ARE AFFECTED BY JAUNDICE

Figure 8 Jaundice

VI. CONCLUSION AND FUTURE WORK

The integration of machine learning into disease management signifies a revolutionary paradigm in healthcare. This technological advancement promises earlier, more accurate diagnoses and personalized treatment plans. Despite existing challenges such as ensuring data privacy and enhancing model robustness, the potential benefits are significant, paving the way for improved patient outcomes and a more efficient healthcare ecosystem. Looking ahead, future enhancements include the exploration of prognosis prediction for disease courses, addressing challenges in healthcare data mining, and refining health policy-making through data-driven insights. This model achieved high accuracy in predicting five specific diseases based on symptom data. The dataset comprised categorical data (1s and 0s representing yes or no), which is well-suited for the random forest algorithm. Cross-validation confirmed the model's effectiveness. However, it's important to note that the model's performance may vary on different datasets with differing diseases, data sizes, or complexities. Future work will investigate the model's generalizability and potential for broader clinical applications. The model's performance in disease prediction can be boosted by focusing on optimizing hyperparameters, exploring feature engineering, and employing ensemble methods. Utilization of cross-validation for robust assessment and careful selection of algorithms are essential. Addressing class imbalances with data augmentation and prioritizing relevant features are crucial steps. Applying regularization to prevent overfitting and integrating domain knowledge for richer insights further enhances performance. The model's performance in disease prediction can be boosted by focusing on optimizing hyperparameters, exploring feature engineering, and employing ensemble methods. Utilization of cross-validation for robust assessment and careful selection of algorithms are essential. Addressing class imbalances with data augmentation and prioritizing relevant features are crucial steps. Applying regularization to prevent overfitting and integrating domain knowledge for richer insights further enhances performance. Healthcare holds the promise of proactive disease outbreak detection, more informed health policy decisions, and enhanced preventive measures, contributing to a healthcare system that is not only responsive but also anticipatory in addressing public health challenges.

REFERENCES

[1] H. E. Massari, Z. Sabouri, S. Mhammedi, and N. Gherabi, “Diabetes Prediction Using Machine Learning Algorithms and Ontology,” Journal of ICT Standardization, vol. 10, no.2, pp. 319-337, 2022.

[2] A. M. Qadri, A. Raza, K. Munir, and M. S. Almutairi, “Effective Feature Engineering Technique for Heart Disease Prediction With Machine Learning,” IEEE Access, vol. 11, pp. 56214-56224, 2023.

[3] Z. Sabouri, Y. Maleh, and N. Gherabi, “Benchmarking Classification Algorithms for Measuring the Performance on Maintainable Applications,” Advances in Information Communication and Cybersecurity, pp. 173-179, 2022.

[4] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar and N. Ullah, “A method for improving prediction of human heart disease using machine learning algorithms,” Mobile Inf. Syst., vol. 2022, pp. 1-9, Mar. 2022.

[5] A. Qayyum, J. Qadir, M. Bilal and A. Al-Fuqaha, “Secure and Robust Machine Learning for Healthcare: A Survey,” IEEE Reviews in Biomedical Engineering, vol. 14, pp. 156-180, 2021

[6] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, “Comparing different supervised machine learning algorithms for disease prediction,” BMC Medical Informatics and Decision Making, pp. 1-16, 2019.

[7] P. S. Kohli and S. Arora, “Application of machine learning in disease prediction,” International Conference on Computing Communication and Automation (ICCCA), pp. 1-4, 2018.

[8] A. Gavhane, G. Kokkula, I. Pandya, and K. Devadkar, “Prediction of heart disease using machine learning,” Second International Conference on Electronics, Communication and Aerospace Technology(ICECA), pp. 1275-1278, 2018.

[9] R. Katarya and P. Srinivas, “Predicting heart disease at early stages using machine learning,” International Conference On Electronics and Sustainable Communication Systems (ICESC), pp. 302-305, 2020.

[10] S. K. Ata, M. Wu, Y. Fang, L. Ou-Yang, C. K. Kwoh, and X. L. Li, Recent advances in network-based methods for disease gene prediction,” Briefings Bioinf., vol. 22, pp. 208-225, 2020.

[11] Y. Hasija, N. Garg, and S. Sourav, “Automated detection of dermatological disorders through image-processing and machine learning,” International Conference on Intelligent Sustainable Systems (ICISS), pp. 1047-105, 2017.

[12] J. Menche A. Sharma, M. Kitsak, S. D. Ghiassian, M. Vidal, J. Loscalzo, and A.-L. Baraba´si, “Uncovering disease-disease relationships through the incomplete interactome,” Science, vol. 347, pp. 450-589, 2015.

[13] A.-L. Baraba´si, N. Gulbahce, and J. Loscalzo, “Network medicine: A network-based approach to human disease,” Nature Rev. Genet., vol. 12, pp. 56-68, 2010.

[14] X. Wang, N. Gulbahce, and H. Yu, “Network-based methods for human disease gene prediction,” Briefings Funct. Genomics, vol. 10, no. 5, pp. 56-68, 2011.

[15] M. M. Ghassemi, T. Naumann, F. Doshi-Velez, et al. “Deep Care: A deep learning approach for multimodal analysis of patient records,” IEEE Journal of Biomedical and Health Informatics, vol:100, pp.55-78, 2019.