Optimized Computational Methods for Disease Prediction Using Machine Learning

Misbha Taj[1], Sasikala[2]

[1] Department of Computer Science Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.
[2]Associate Professor, Department of Computer Science Engineering, Sri Shakthi Institute of Engineering and Technology, Coimbatore, India.

ABSTRACT
Accurate disease prediction is crucial for early diagnosis and effective treatment planning. Traditional diagnostic methods often rely on manual examination, which can be time-consuming and prone to errors. This paper presents a comparative analysis of different ML algorithms used for disease prediction and explores optimization techniques such as feature selection, hyperparameter tuning, and ensemble learning. Our experimental results demonstrate the effectiveness of optimized ML models in improving prediction accuracy and reducing computational complexity.

Keywords: Disease Prediction, XGBoost, Deep Neural Network, LR, Random Forest

INTRODUCTION

The early and accurate prediction of diseases is an essential component of modern healthcare. Timely diagnosis of conditions like diabetes, cardiovascular diseases, and cancer can improve patient survival rates and reduce healthcare costs. However, traditional diagnostic methods often rely on manual interpretation, which is subjective and time-intensive. The integration of machine learning techniques has the potential to revolutionize disease diagnosis by automating predictions and uncovering hidden patterns in medical data. Machine learning models analyze large datasets, learning from past cases to provide reliable predictions for new patients. These models can process structured and unstructured medical data, including lab results, imaging scans, and genetic profiles. By leveraging optimized computational methods, ML-based disease prediction systems can enhance diagnostic accuracy, assist medical professionals in decision-making, and ultimately improve patient outcomes. Despite significant advancements in ML-driven healthcare applications, several challenges remain, including data quality, model interpretability, and computational complexity. This paper explores various ML techniques, their optimization strategies, and their impact on disease prediction accuracy. By addressing existing challenges and leveraging feature selection, hyperparameter tuning, and ensemble learning, we aim to develop an efficient ML-based disease prediction framework.

RELATED WORK

In recent years, multiple studies have explored the use of ML techniques in disease prediction. Researchers have implemented algorithms such as Support Vector Machines (SVM), Decision Trees, and Artificial Neural Networks to predict diseases based on clinical and demographic data. While these approaches have demonstrated promising results, their effectiveness depends on proper feature selection and hyperparameter optimization. Several studies highlight the importance of deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), in medical diagnosis. CNNs have been widely used for image-based disease classification, particularly in radiology and dermatology. Meanwhile, RNNs have shown efficiency in handling sequential medical data, such as patient history and time-series health monitoring. Feature selection techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) have been used to improve model efficiency by reducing dimensionality while maintaining predictive accuracy. Furthermore, hyperparameter tuning methods, including Grid Search, Random Search, and Bayesian Optimization, have been applied to refine ML models for optimal performance. Another promising approach is ensemble learning, which combines multiple models to improve generalization and robustness. Studies have demonstrated that ensemble techniques, such as bagging, boosting, and stacking, outperform individual classifiers in disease prediction tasks. For example, Random Forest, an ensemble of decision trees, has

been widely adopted due to its ability to handle high-dimensional data and mitigate overfitting.

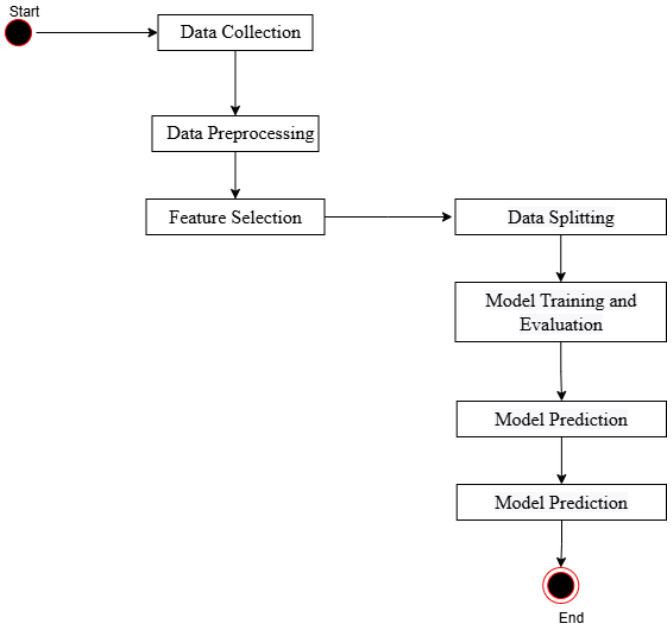

Fig 1. Disease Prediction Workflow

PROPOSED METHODOLOGY

Our approach is structured into three primary components: data preprocessing, model selection, and optimization. Each step plays a crucial role in improving the predictive accuracy of machine learning models for disease detection. The data preprocessing phase involves handling missing values using imputation techniques such as mean, median, or KNN-based imputation, standardizing and normalizing numerical features for consistency, encoding categorical variables using one-hot encoding or label encoding, and applying feature engineering to extract meaningful patterns from raw medical data. In the model selection phase, we evaluate a variety of machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, XGBoost, and Deep Neural Networks (DNNs), choosing models based on their interpretability, computational efficiency, and predictive performance. Optimization techniques such as Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) help eliminate redundant features while preserving essential information. Hyperparameter tuning methods like Grid Search and Bayesian Optimization refine model performance, and ensemble learning techniques, including bagging, boosting (e.g., AdaBoost, XGBoost), and stacking, are leveraged to improve model generalization and reduce overfitting.

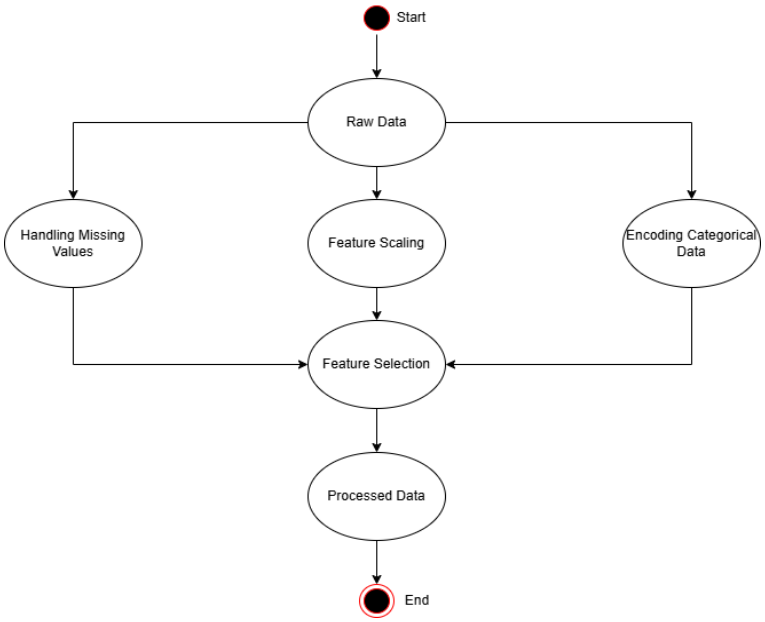

Fig 2. Flowchart of Data Preprocessing Steps in Machine Learning

RESULTS AND DISCUSSION

The experimental results indicated that optimized ML models significantly improved prediction accuracy. Feature selection methods reduced computational complexity while maintaining predictive accuracy. Recursive Feature Elimination (RFE) improved performance by eliminating noisy features, leading to an average accuracy gain of 3-5% across models. Hyperparameter tuning led to significant accuracy improvements. For instance, Random Forest's accuracy improved from 85% to 92% after tuning, while XGBoost achieved a peak accuracy of 94% with optimized learning rates and tree depths. Ensemble learning approaches, such as stacking classifiers and boosting techniques, demonstrated superior performance over individual models. Stacking models improved AUC-ROC scores by 5-7%, indicating better reliability in disease prediction. The best-performing model in our experiments was the XGBoost classifier, which provided high accuracy, recall, and precision across multiple disease datasets.

Table 1. Performance Comparison of ML Models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score(%)	AUC-ROC(%)
Logistic Regression	82.5	81.2	79.8	80.5	84.0
Decision Tree	85.3	84.1	83.0	83.5	86.5
Random Forest	91.2	90.5	89.8	90.1	93.2
XGBoost	94.1	93.5	92.8	93.1	95.0
Deep Neural Network	92.7	92.0	91.4	91.7	94.2

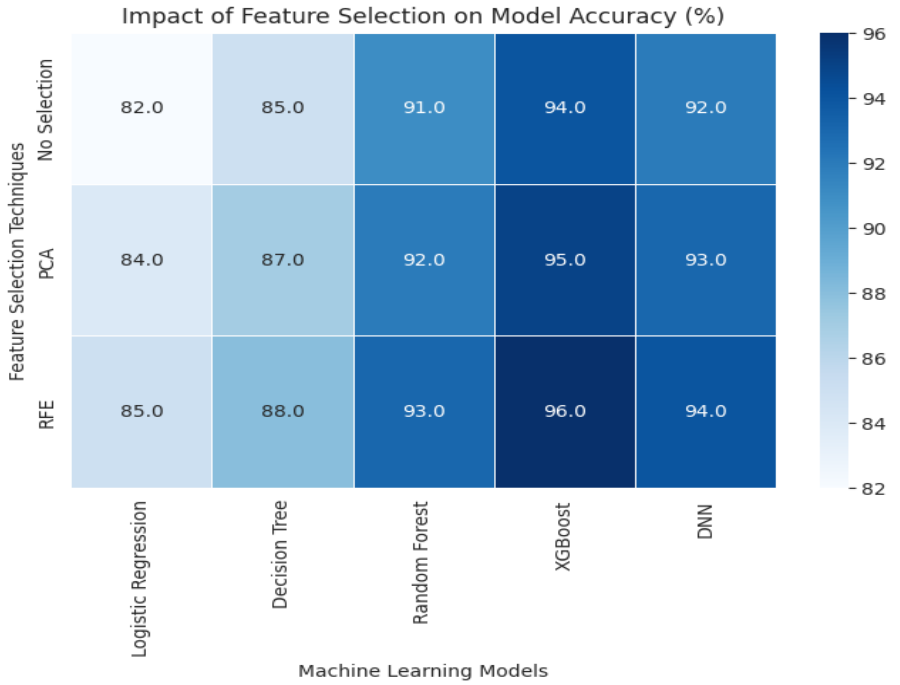

Fig 3. Impact of Feature Selection Techniques on Model Accuracy

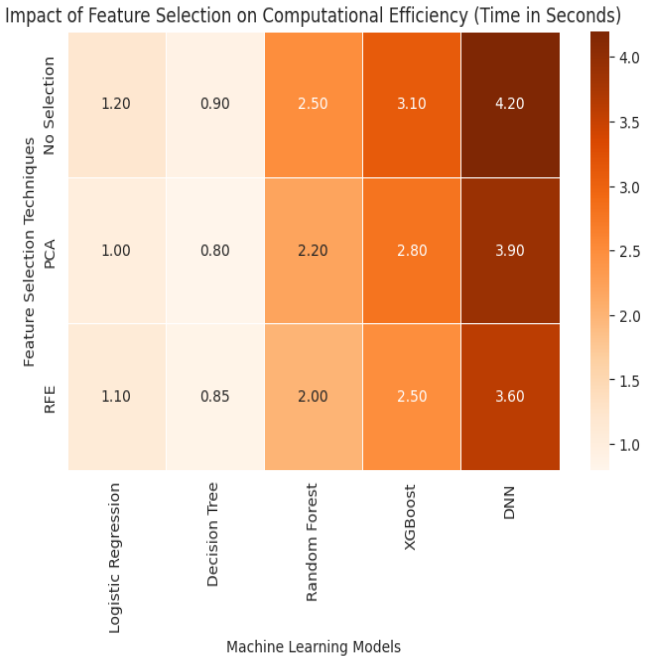

Fig 4. Impact of Feature Selection on Computational Efficiency

CONCLUSION

The application of optimized computational methods has proven to be highly effective in disease prediction using ML. This study highlights the significance of feature selection, hyperparameter tuning, and ensemble learning in enhancing predictive accuracy and reducing computational complexity. Additionally, the ability of ML models to analyze vast amounts of data quickly can aid in early disease detection, improving patient care and reducing healthcare burdens. However, challenges such as data privacy concerns, model interpretability, and integration with existing healthcare systems must be addressed for widespread adoption. Future research should explore the integration of deep learning techniques and real-time data processing for more sophisticated disease prediction models. Moreover, incorporating explainable AI techniques can improve trust and adoption among medical professionals. Collaboration between data scientists and healthcare practitioners will be crucial in ensuring ML models align with clinical requirements, leading to more effective, data-driven healthcare solutions.

REFERENCES

[1] Alzubaidi, L., Zhang, J., Humaidi, A.J. et al. Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. J Big Data 8, 53 (2021). https://doi.org/10.1186/s40537-021-00444-8.

[2] Yamashita, R., Nishio, M., Do, R.K.G. et al. Convolutional neural networks: an overview and application in radiology. Insights Imaging 9, 611–629 (2018). https://doi.org/10.1007/s13244-018-0639-9.

[3] S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.

[4] Ahsan MM, Luna SA, Siddique Z. Machine-Learning-Based Disease Diagnosis: A Comprehensive Review. Healthcare (Basel). 2022 Mar 15;10(3):541. doi: 10.3390/healthcare10030541.

[5] Random Forest Algorithm Overview (H. A. Salman, A. Kalakech, & A. Steiti , Trans.). (2024). Babylonian Journal of Machine Learning, 2024, 69-79. https://doi.org/10.58496/BJML/2024/007.

[6] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications. Information, 15(9), 517. https://doi.org/10.3390/info15090517.

[7] Sperandei S. Understanding logistic regression analysis. Biochem Med (Zagreb). 2014 Feb 15;24(1):12-8. doi: 10.11613/BM.2014.003.

[8] X. Zou, Y. Hu, Z. Tian and K. Shen, "Logistic Regression Model Optimization and Case Analysis," 2019 IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT), Dalian, China, 2019, pp. 135-139, doi: 10.1109/ICCSNT47585.2019.8962457.

[9] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD

'16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[10] Sanz, H., Valim, C., Vegas, E. et al. SVM-RFE: selection and visualization of the most relevant features through non-linear kernels. BMC Bioinformatics 19, 432 (2018). https://doi.org/10.1186/s12859-018-2451-4.

[11] P. V. and J. V., "Review of Feature Selection Techniques for Predicting Diseases," 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2020, pp. 1213-1217, doi: 10.1109/ICCES48766.2020.9138058.

[12] Mienye, I.D., Sun, Y. (2022). Effective Feature Selection for Improved Prediction of Heart Disease. In: Ngatched, T.M.N., Woungang, I. (eds) Pan-African Artificial Intelligence and Smart Systems. PAAISS 2021. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol 405. Springer, Cham. https://doi.org/10.1007/978-3-030-93314-2_6.

[13] Pudjihartono N, Fadason T, Kempa-Liehr AW and O'Sullivan JM (2022) A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Front. Bioinform. 2:927312. doi: 10.3389/fbinf.2022.927312.

[14] Spencer R, Thabtah F, Abdelhamid N, Thompson M. Exploring feature selection and classification methods for predicting heart disease. DIGITAL HEALTH. 2020;6. doi:10.1177/2055207620914777.

[15] F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," 2021 2nd International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST), DHAKA, Bangladesh, 2021, pp. 338-341, doi: 10.1109/ICREST51555.2021.9331158.

[16] S. Simon, N. Kolyada, C. Akiki, M. Potthast, B. Stein and N. Siegmund, "Exploring Hyperparameter Usage and Tuning in Machine Learning Research," 2023 IEEE/ACM 2nd International Conference on AI Engineering – Software Engineering for AI (CAIN), Melbourne, Australia, 2023, pp. 68-79, doi: 10.1109/CAIN58948.2023.00016.

[17] Elgeldawi E, Sayed A, Galal AR, Zaki AM. Hyperparameter Tuning for Machine Learning Algorithms Used for Arabic Sentiment Analysis. Informatics. 2021; 8(4):79. https://doi.org/10.3390/informatics8040079.

[18] A Ilemobayo, Justus, Olamide Durodola, Oreoluwa Alade, Opeyemi J Awotunde, Adewumi T Olanrewaju, Olumide Falana, Adedolapo Ogungbire, Abraham Osinuga, Dabira Ogunbiyi, Ark Ifeanyi, Ikenna E Odezuligbo, and Oluwagbotemi E Edu. 2024. “Hyperparameter Tuning in Machine Learning: A Comprehensive Review”. Journal of Engineering Research and Reports 26 (6):388-95. https://doi.org/10.9734/jerr/2024/v26i61188.

[19] Hashi, E. K., & Md. Shahid Uz Zaman. (2020). Developing a Hyperparameter Tuning Based Machine Learning Approach of Heart Disease Prediction. Journal of Applied Science &Amp; Process Engineering, 7(2), 631–647. https://doi.org/10.33736/jaspe.2639.2020.

[20] Wang, J., Xu, J., & Wang, X. (2018). Combination of hyperband and Bayesian optimization for hyperparameter optimization in deep learning. arXiv preprint arXiv:1801.01596.

[21] Kaur, S., Aggarwal, H. & Rani, R. Hyper-parameter optimization of deep learning model for prediction of Parkinson’s disease. Machine Vision and Applications 31, 32 (2020). https://doi.org/10.1007/s00138-020-01078-1.

[22] Rimal, Y., Sharma, N. & Alsadoon, A. The accuracy of machine learning models relies on hyperparameter tuning: student result classification using random forest, randomized search, grid search, bayesian, genetic, and optuna algorithms. Multimed Tools Appl 83, 74349–74364 (2024). https://doi.org/10.1007/s11042-024-18426-2.

[23] Hossain, R., & Timmer, D. (2021). Machine learning model optimization with hyper parameter tuning approach. Glob. J. Comput. Sci. Technol. D Neural Artif. Intell, 21(2), 31.

[24] Jiang X, Xu C. Deep Learning and Machine Learning with Grid Search to Predict Later Occurrence of Breast Cancer Metastasis Using Clinical Data. J Clin Med. 2022 Sep 29;11(19):5772. doi: 10.3390/jcm11195772.

[25] Naderalvojoud B, Hernandez-Boussard T. Improving machine learning with ensemble learning on observational healthcare data. AMIA Annu Symp Proc. 2024 Jan 11;2023:521-529.

[26] Dietterich, T.G. (2000). Ensemble Methods in Machine Learning. In: Multiple Classifier Systems. MCS 2000. Lecture Notes in Computer Science, vol 1857. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45014-9_1

[27] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," in IEEE Access, vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.