iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 12 **Issue:** XI **Month of publication:** November 2024

DOI: https://doi.org/10.22214/ijraset.2024.65223

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

Extensive Review on Multiple Disease Detection using Machine Learning

Geetanjali Bansod[1], Parul Patil[2], Yash Patel[3], Prajakta Jagtap[4]
Dept. of Computer, KJ College of Engineering and Management Research Pune, India

Abstract: ***Diagnosing diseases rapidly are critical for improving patient outcomes, yet the usual diagnostic pathway causes significant delays due to extended waiting periods between sessions and advanced diagnostic tools. This review paper presents a study on the application of machine learning algorithms for multi-disease identification utilizing crowdsourced symptoms and a web-based platform. In this study, we want to investigate how machine learning algorithms might enhance diagnosis speed and accuracy by predicting diseases based on symptom data. Integrating such a technology into practice workflows can assist healthcare providers in prioritizing testing procedures based on detected risk, allowing for quicker patient diagnosis and treatment. Existing literature on machine learning applications in disease detection Challenges prevailing today for multi-disease prediction currently covered within the review. In the end, this application potentially stands to improve patient care access by eliminating complex cases and thereby freeing up clinical expert time for more difficult patients as well as data-driven insights that can lead to better healthcare outcomes.***

Keywords: Machine Learning Algorithms, Evaluation Metrics, Supervised Learning, Electronic health records

I. INTRODUCTION

Predicting and early detection of diseases have, over the years been a vital component of healthcare because timely interventions can notably reduce morbidity and mortality. Traditionally, healthcare professionals relied mainly on patient history, physical examination, and other diagnostic tests that could be conducted on the patient to diagnose diseases. However, the volume of medical data and, more crucially, the complexity of many of today's diseases have increasingly established a strong need for methods of prediction and diagnosis to be more sophisticated than they have been in the past. Machine learning (ML) is revolutionizing healthcare by enabling significant advancements. A Multiple Disease Prediction System (MDPS) is an advanced computational approach that leverages machine learning algorithms to assess an individual's likelihood of developing certain diseases based on medical data. Such systems always seek early detection of Alzheimer's disease, heart disease, diabetes, cancer, and so forth at an appropriate cost value, much before it manifests itself in terms of symptoms. Thus, such a tool is highly useful in the comprehensive monitoring of health if MDPS can simultaneously predict more than one disease. MDPS is a proposal involving machine learning in transforming the healthcare sector in predictive analysis of various diseases from clinical data. MDPS, being an algorithm, will analyze patient data for complex patterns, spot early signs in diseases, warn/treat in time, and allow preventive measures to seek for treatment purposes. The essence of the MDPS is in forecasting these diseases into a single system, which helps enhance patient care, minimize the resource burden in healthcare, and give much-needed relief from the huge infrastructural demands to contain disease spread.

The design of multiple disease prediction systems can be traced back to the classical applications of AI to medicine in the 1960s and 1970s. The pioneers had been expert systems, such as MYCIN and INTERNIST-I, facing the wilderness of diseases affecting associated diagnosis. These systems were truly humane at their approach but were limited in their applications due to the extremely narrow realm of diseases they could cater to, rigid rule-based frameworks, and an inability to adapt to varied situations in medicine. When data-driven, machine learning algorithms appeared in the 1990s, predictive modeling took a quantum leap towards changing the shape and everything, bringing in mandatory efforts in favor of precision medicine. Dynamic disease prediction was thus made possible through other algorithms like decision trees, SVM, further allowing neural networks to make disease prediction models more mobile and data-centered. It x-rayed the manual-based programming rules that governed various disease prediction models that induced learning and modeled historical behaviors to make predictions about disease progression. That said, the original models were mostly disease specific; they targeted chronic conditions like heart disease, diabetes, and cancer individually, as opposed to collectively dealing with those conditions using a single system approach. The beginning growth of MDPS started to be apparent in the 2000s, where developments in both computational power and the multiplexing between big data and medical imaging gave push to machine learning applications within healthcare.

Moreover, with the incorporation of Electronic Health Records (EHRs), vast amounts of retrospective patient data can feed models that take into consideration a broader scope of variables in predicting various conditions at the same time. The integration of ensemble methods, multi-label classification, and advanced [1]deep learning approaches, such as [1]convolutional and recurrent [1]neural networks, has enhanced prediction accuracy and broadened the capabilities of systems designed to address multiple diseases. Currently, Multiple Disease Prediction Systems (MDPS) utilize state-of-the-art machine learning techniques, including deep learning, natural language processing (NLP), and federated learning, to predict various diseases like cardiovascular conditions, Alzheimer's disease, diabetes, and cancer with greater precision. Providing real-time data analysis, wearable device integration, and data storage in the cloud, modern MDPS can offer continuous monitoring with personalized health insights, thus advancing traditional healthcare diagnostics.

A. Ensemble Learning Approaches

Various studies have utilized ensemble learning approaches to tackle the issue of multiple disease prediction accuracy. For instance, Zhang et al. (2020) suggested an ensemble model combining decision trees and SVMs for predicting cardiovascular diseases, diabetes, and hypertension using patient data. Their model turned out to have high accuracy, thereby controlling overfitting by means of cross-validation.

B. Deep Learning Models

Prayer Great opposes have been called and admitted by the report, which the MDPS seem involving deep learning as a major tool especially handling complex datasets like the images and misshaped data: Convolutional Neural Networks (CNNs) were used by Gonzalez et al. (2020) to evaluate medical imaging data for illness identification, with an emphasis on early-stage cancer diagnosis. High precision and notable advancements over conventional techniques were offered by their method.

The application of Long Short-Term Memory (LSTM) networks was explored to predict the progression of chronic diseases using sequential patient data. Their method effectively captured temporal relationships, showing promising outcomes in predicting conditions such as diabetes and cardiovascular disease.

C. Methods for Classifying Multiple Labels

It is worth noting that multi-label classification methods were used because one patient could have multiple disorders at the same time: Kumar et al. (2019) created a multi-label classification framework using SVM and K- Nearest Neighbors to predict a variety of diseases from clinical data. According to their findings, multi-label approaches have the potential to greatly increase disease detection rates.

D. Feature Selection and Dimensionality Reduction -

This is utmost critical for betterment of model performance and interpretability: Sharma et al. (2020) explored different feature selection techniques used to improve disease predictions in multivariate data sets. Techniques such as Recursive Feature Elimination and Lasso regression were applied to identify important cardiovascular disease risk factors with minimized overfitting.

E. Explainable AI in MDPS

The openly acknowledged need for transparency in ML models has resulted in the drive to incorporate explainable AI techniques in MDPS: Khan et al. (2021) provided an AI explaining framework for refuting or saying what could be predicted by their machine learning models with respect to multiple diseases. They used the SHAP values to throw light upon the importance of features and clarify their predictions for healthcare practitioners.

II. PROBLEM FINDING

It is yet another formidable and excellent question of developing a multiple disease prediction system that can fairly diagnose heart disease and diabetes with the help of machine learning. Several fundamental issues and challenges were encountered during the research, maturation, and establishment of such systems. They are:

A. Missing Values

Many datasets are afflicted by incomplete records due to a multitude of reasons, like patients not undergoing certain tests or incomplete data gathering in the first place. Machine learning algorithms tend to repeat the same error pattern with the given dataset,

which means they can bias predictions based on incomplete or wrong input. There are quite a few ways to restore missing values by sourcing values that take into account affected records. Some of the ways include filling with average values or utilizing more sophisticated methods, such as k-nearest neighbor (KNN) imputation; nonetheless, these strategies do not always work well.

B. Noisy Data

Medical data is sometimes termed as noisy, meaning it contains measurement errors or outliers, which seriously affect the performance of a prediction model. For instance, while the results of lab tests might be affected by measurement errors, those results could also be subject to other inconsistencies.

C. Class imbalance

In formulations pertaining to the healthcare dataset, some diseases are far rarer in contrast to others. For instance, while one dataset might contain heart disease records and diabetes, the positive examples of heart disease or diabetes might be significantly outnumbered by the negative cases. These kinds of machine learning models trained on data with inherent class imbalance could easily favor the majority class - in this case healthy patients - which leads to a high accuracy score but poor prediction for the diseased persons. This requires balancing techniques like oversampling (like Synthetic Minority Over-Sampling Technique, SMOTE) or under sampling, but these balancing techniques need to be careful in their application since, otherwise, they can introduce bias.

D. Complexity of Black-Box Models

On the other hand, deep learning, which may enhance prediction accuracy, is extremely interpreted. Unless doctors are satisfied with how the specific input features (age of the patient, blood pressure, et cetera) predict the heart disease or make a prediction of high risk for heart disease, they may not be inclined to take the classes or those clinical models.

E. Common VS Unique Predictors

The prediction for diseases such as obesity is common, while some others are more specific. Thus, it becomes a problem ensuring the model does not mix up these shared and unique predictors without losing function. It must learn to discriminate when X (like high cholesterol) is contributing more to (the risk of) heart disease than diabetes and vice versa.

F. Curse of Dimensionality

Instead, too much data can lead to overfitting, which occurs when a model can partially memorize and perform well based on training data but then perform badly on fresh, unknown data. This allows the model to perform better on the training dataset than it would on a new dataset. Obtaining shaded features for both disorders is not easy and may necessitate approaches such as correlation analysis, or principal component analysis (PCA), or recursive feature elimination (RFE).

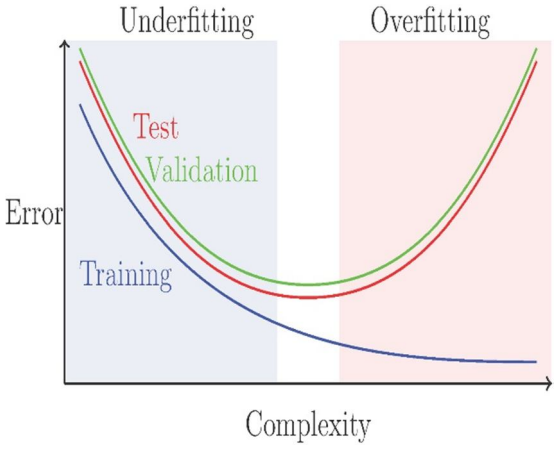


Fig. 1: Bias-Variance Tradeoff

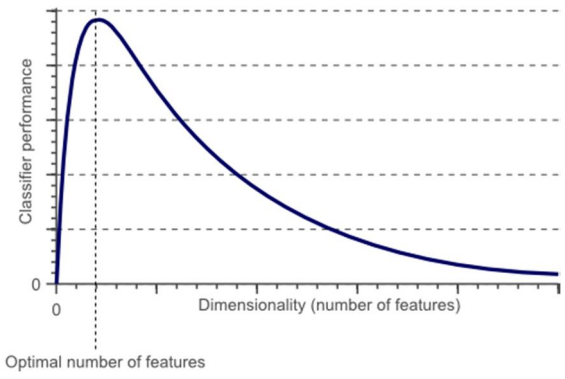


Fig. 2: Dimensionality

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 12 Issue XI Nov 2024- Available at www.ijraset.com

III. GENERAL METHODOLOGY

This section provides a detailed explanation of the dataset creation process, model development, and disease prediction approach. Data collection serves as the first step, where our system gathers both labelled and unlabeled data from various sources. After data collection, preprocessing is conducted, dividing the data into training and testing sets.[2] The training data is then applied to machine learning algorithms, like CNN and KNN, over multiple epochs to improve prediction accuracy. Once the model reaches the desired accuracy after repeated epochs, it is prepared for testing.

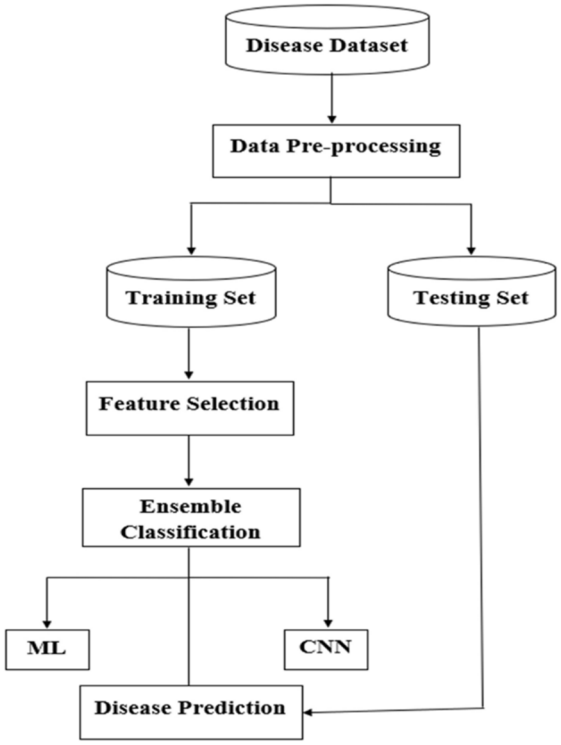


Fig. 3: Flow diagram of MDPS

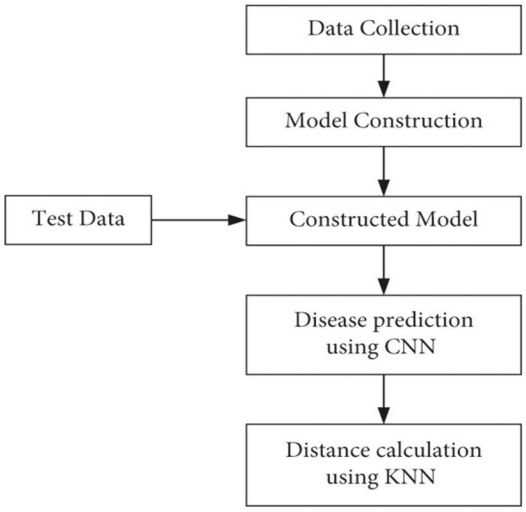


Fig. 4: Steps for Training Model

[3]The model is subsequently evaluated using a test dataset containing new data that was not part of the training process. This step enables us to measure the model's performance on unfamiliar data, as demonstrated in Figs. 1 and 2. If the model meets the required accuracy during testing, it is ready for deployment, as illustrated.

A. Evaluation Metrics

The disease prediction model is assessed using performance metrics. The confusion matrix consists of true positives (TP), representing correct predictions of individuals with chronic diseases; [4]true negatives (TN), representing correct predictions of healthy individuals; false positives (FP), where healthy individuals are incorrectly predicted as diseased; and false negatives (FN), where diseased individuals are incorrectly predicted as healthy. The four performance evaluation metrics are outlined as follows:

- **Accuracy**

Classification accuracy is defined as the ratio of correctly predicted values to the total number of predictions, and it is mathematically expressed as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} * 100. \quad \ldots\ldots (1)$$

- **Precision**

Precision, also known as positive predictive value (PPV), is defined as the ratio of correct predictions to the total number of positive predictions, including both true and false positives. It is mathematically represented as follows:

$$\text{Precision} = \frac{TP}{TP + FP}. \quad \ldots\ldots (2)$$

- Recall

Recall, also known as sensitivity or true positive rate (TPR), is defined as the ratio of correctly predicted positive values to the sum of true positive predictions and false negatives. It is mathematically represented as follows:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad \ldots\ldots (3)$$

- F1-Score

Additionally, the F1 − Score value is quite appropriate whenever the false positive and negative values differ. This is how the F1 − Score is represented mathematically:

$$F_\beta = \frac{\left(1 + \beta^2\right)(\text{Precision} * \text{Recall})}{\left(\beta^2 * (\text{Precision} + \text{Recall})\right)}. \quad \ldots\ldots (4)$$

Through simplification, β = 1.

$$F_1 - \text{Score} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}. \quad \ldots\ldots (5)$$

The precision, F1-score, and recall of the proposed CNN and KNN model are compared to those of the Naïve Bayes, Logistic Regression models, and, Decision Tree as summarized in the table. Accuracy is highlighted as the most important metric since the prediction outcome is critical for the patient; an inaccurate prediction could adversely affect them. The model's performance is also evaluated using additional metrics such as F1-score, precision, and recall.

Figure 5 presents the accuracy levels of four machine learning algorithms where- Naïve Bayes shows the lowest accuracy among the four models, Decision Tree has a moderate accuracy, slightly higher than Naïve Bayes, logistic Regression achieves a higher accuracy than both Naïve Bayes and Decision Tree and CNN & KNN has the highest accuracy, surpassing all other models.

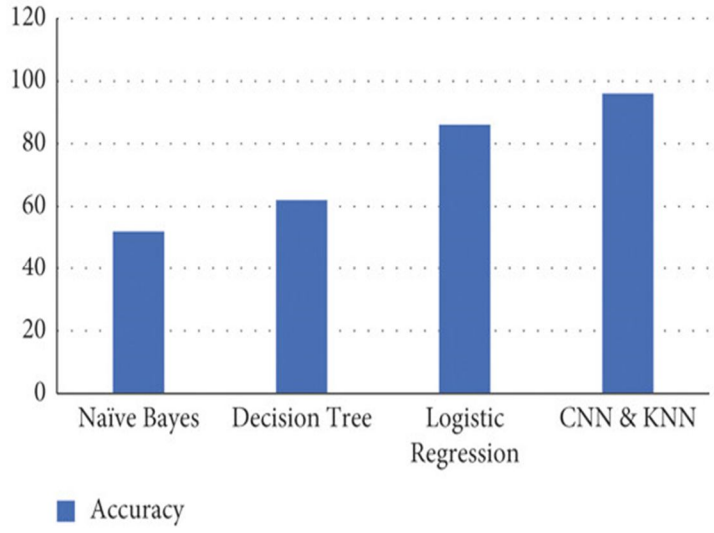


Fig. 5: Accuracy for ML algorithms in MDPS

The figure 6 shows a comparison of four machine learning models- Naïve Bayes shows the lowest precision and relatively lower recall and F1-score, Decision Tree has moderate values for all three metrics, with similar scores for[2] precision, recall, and F1-score, Logistic Regression achieves higher performance, with closely aligned values for all metrics, indicating good balance, and CNN & KNN demonstrates the highest performance across F1-score, recall, and precision, indicating that this combined approach is the most effective among the models compared.

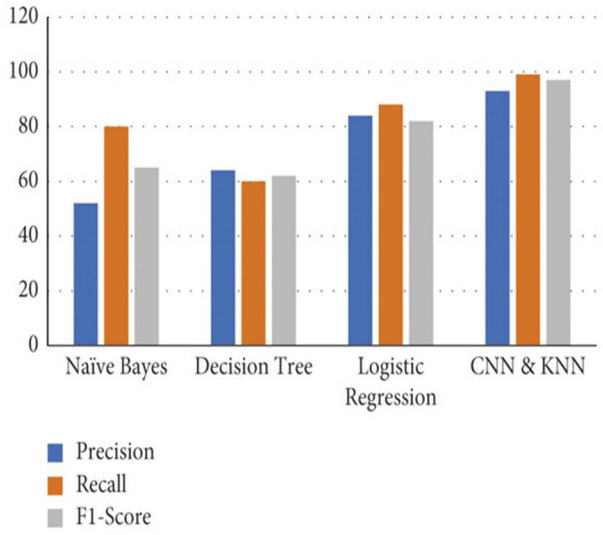


Fig. 6: Precision, Recall, F1- score for ML Algorithms in MDPS

IV. EXPECTED SOLUTIONS

This work introduces the Multiple Disease Prediction System (MDPS), a set of machine learning algorithms for predicting both cardiovascular disease and diabetes. The proposed system integrates data from system sources such as medical histories, lab tests, and demographic data with a combination of classifiers trained on both diseases. The proposed process is in the following consecutive steps:

1) *Data Preparation:* Missing value imputation, data normalization, and balancing of dataset will be done applying such techniques as Synthetic Minority Over-sampling Technique (SMOTE).
2) *Feature Selection:* Finding the most pertinent features for diabetes and heart disease requires the use of recursive feature elimination (RFE) and correlation analysis.
3) *Model Training:* The system will employ a hybrid ensemble method that includes decision trees, support vector machines, and random forests. Model will be trained separately on heart disease and diabetes datasets, and prediction will be arrived at from each dataset using majority voting.
4) *Evaluation:* The model will be validated using cross-validation and accuracy, precision, recall, and F1 scores will be calculated. Individual predictions will be explained by increasing model interpretability through the use of SHAP (Shapley Additive Explanations) values.

V. RESULTS AND DISCUSSIONS

The proposed Multiple Diseases Prediction System (MDPS) was created to conduct testing on minor computations using various benchmark datasets, including the datasets like Cleveland Heart Disease and the Pima Indians Diabetes. Using a combination of Decision Trees, Random Forests, and Support Vector Machines (mixed ensemble), along with certain data preprocessing techniques such data imputation, the classification models were produced and class balance with SMOTE. These gave their predictions as follows:

Heart Disease Prediction Accuracy: The ensemble model's accuracy in predicting heart disease was 88%, with a precision measure of 0.9%, recall of 82%, and F1 score of 86%.

Diabetes Prediction Accuracy: The system got a successful accuracy of 85 with precision 84, recall:80, F1 score of 86.

Multi-task Learning: The model uniquely performed quite competently in predicting both diseases simultaneously, only slightly less efficient than the case of predicting each disease independently. SHAP values explain how each of the features has contributed to the predictions, thereby increasing the transparency of the model.

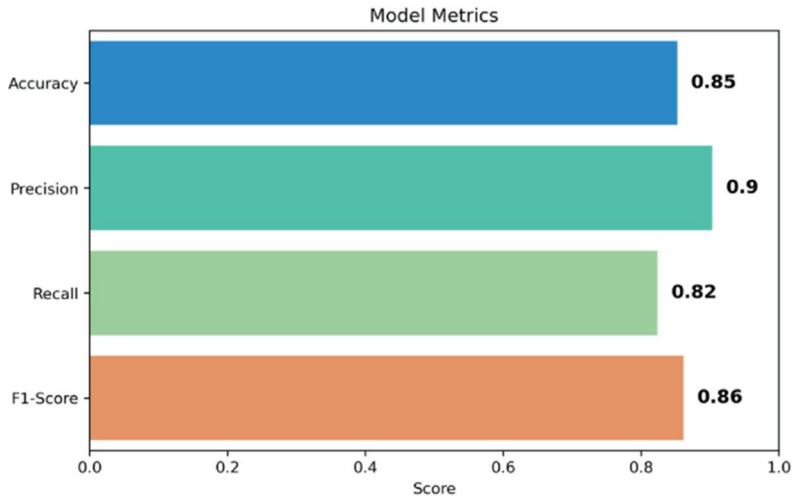


Fig. 7: Evaluation metrics for MDPS

VI. CONCLUSION & FUTURE SCOPE

This paper explores various ways in which machine learning could have previously enhanced the detection of six diseases, positing that advancements in machine learning are essential for driving significant improvements in healthcare. By analyzing complex patterns in patient data, [1]machine learning models like Decision Trees, Support Vector Machines (SVMs), Random Forests, and Neural Networks have significantly increased diagnostic accuracy. Problems like data quality, feature selection, are yet some other problems to consider, requiring further research. Finally, web-based solutions are of paramount importance and allow timely disease detection by providing clinical experts with instantaneous access to diagnostic instruments, reducing patient wait time and increasing efficiency in healthcare. The present project thus intervenes in this area by developing a web application that predicts multiple diseases based on symptom information volunteered by the user. The tool carries out an assessment of symptoms during the first stage, thus easing the workload on the healthcare professionals by giving them insight from the data before sending the patient for confirmation of the disease. Moreover, the integration of machine learning systems with the increase of web-based solutions will keep playing a great role in enhancing the detection and management of diseases in the future.

There is immense future potential for a machine-learning-based MDPS to, along with personalized health care, take the healthcare society onto the next level. The other research areas and improvements that are yet to be undertaken include:

1.Incorporating More Diseases: The MDPS might become a more comprehensive tool for health risk assessment by extending prediction to other chronic illnesses like Alzheimer's disease, stroke, or cancer.

2. Use of Deep Learning: The specificity of prediction would be enhanced using a deep learning approach like CNNs or RNNs on real-life medical data. These networks particularly outperform traditional algorithms when combined with time series data (e.g., patient health progression).

3. Integrating Real Clinical Data: Future work should focus on incorporating patient's real-time data from electronic health record systems and wearable devices. These initiatives would assist clinical setting applicability and significantly increase predictive competency.4. Personalized Medicine: The system will be more personalized if made in such a way that predictions are based upon genetic data, lifestyle, and social determinants of health. This enables tailored interventions for the individual patient.

5. Model Validation and Clinical Trials: Developmental efforts in this regard would need to be geared towards establishing it amongst the clinicians and patients, including successful testing in clinical trials within real-world setups, to establish its effectiveness for regulatory clearance.

6. Improved Explainability: Future work should zero in on the techniques of explainable AI (XAI) so that as the opacity of complex models passes away, the transparency in model decisions communicates the predictions clearly put forward to patient caregivers.

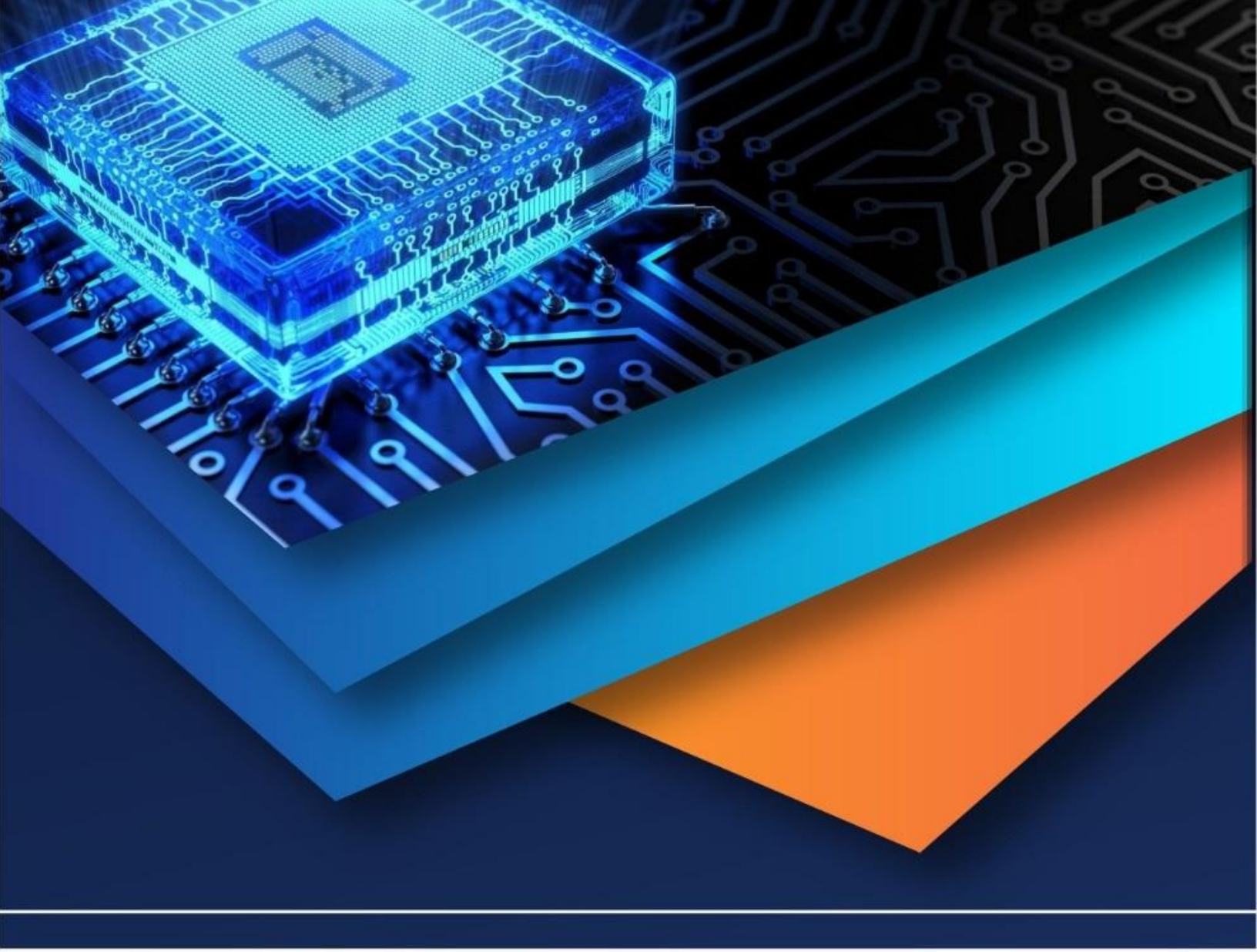
REFERENCES

[1] S.J. Xavier Savarimuthu, Sivakannan Subramani, Alex Noel Joseph Raj. "Artificial Intelligence for Multimedia Information Processing - Tools and Applications", CRC Press, 2024

[2] Shubhangi S. Shambharkar, Pradnya S. Moon, Prachi A. Bainalwar, Shubhangi M. Boarkar. "Machine Learning-Based Approach for Early Detection and Prediction of Chronic Diseases" , 2023 1st DMIHER International Conference on Artificial Intelligence in Education and Industry 4.0 (IDICAIEI), 2023

[3] Saravanan Krishnan, Ramesh Kesavan, B. Surendiran, G. S. Mahalakshmi. "Handbook of 6

[4] David Walker, Magdalena Smigaj, Nebo Jovanovic. "Ephemeral sand river flow detection using satellite optical remote sensing" , Journal of Arid Environments, 2019
[5] DESHMUKH, Sunita P. et al. Hybrid Deep Learning Method for Detection of Liver Cancer. Computer Assisted Methods in Engineering and Science, [S.l.], v. 30, n. 2, p. 151–165, mar. 2023. ISSN 2956-5839.
[6] L. D. Gopisetti, S. K. L. Kummera, S. R. Pattamsetti, S. Kuna, N. Parsi, and H. P. Kodali, "Multiple Disease Prediction System using Machine Learning and Streamlit," Proceedings - 5th International Conference on Smart Systems and Inventive Technology, ICSSIT 2023, pp. 923–931, 2023, doi: 10.1109/ICSSIT55814.2023.10060903.
[7] M. Alsulmi and R. Alshamarani, "Framework for tasks suggestion on web search based on unsupervised learning techniques," Journal of King Saud University - Computer and Information Sciences, vol. 34, no. 8, pp. 5525–5532, Sep. 2022, doi: 10.1016/J.JKSUCI.2021.06.004.
[8] R. Alanazi, "Identification and Prediction of Chronic Diseases Using Machine Learning Approach," J Healthc Eng, vol. 2022, no. 1, p. 2826127, Jan. 2022, doi: 10.1155/2022/2826127.
[9] R. Keniya et al., "Disease Prediction From Various Symptoms Using Machine Learning," SSRN Electronic Journal, Jul. 2020, doi: 10.2139/SSRN.3661426.
[10] K. Arumugam, M. Naved, P. P. Shinde, O. Leiva-Chauca, A. Huaman-Osorio, and T. Gonzales-Yanac, "Multiple disease prediction using Machine learning algorithms," Mater Today Proc, vol. 80, pp. 3682–3685, Jan. 2023, doi: 10.1016/J.MATPR.2021.07.361.
[11] I. Mohit, K. S. Kumar, A. U. K. Reddy, and B. S. Kumar, "An Approach to detect multiple diseases using machine learning algorithm," J Phys Conf Ser, vol. 2089, no. 1, p. 012009, Nov. 2021, doi: 10.1088/1742-6596/2089/1/012009.
[12] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Comput Sci, vol. 165, pp. 292–299, Jan. 2019, doi: 10.1016/J.PROCS.2020.01.047.
[13] N. Ahmed et al., "Machine learning based diabetes prediction and development of smart web application," International Journal of Cognitive Computing in Engineering, vol. 2, pp. 229–241, Jun. 2021, doi: 10.1016/J.IJCCE.2021.12.001.
[14] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," IEEE Access, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.
[15] Z. Wang, J. W. Chung, X. Jiang, Y. Cui, M. Wang, and A. Zheng, "Machine Learning-Based Prediction System For Chronic Kidney Disease Using Associative Classification Technique," International Journal of Engineering & Technology, pp. 1161–1167, 2018, Accessed: Oct. 21, 2024. [Online]. Available: www.sciencepubco.com/index.php/IJET



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY