Leveraging Machine Learning for Health Recognition

Rahul Shoundic[1]
Computer Science, KIET group of institutions, Ghaziabad, India. rahulshoundic@gmail.com

Saumya Maheshwari[2]
Computer Science, KIET group of institutions, Ghaziabad, India.
saumyamaheshwari20122003@gmail.com

Vivek Yadav[3]
Computer Science, KIET group of institutions, Ghaziabad, India. vivo16012001@gmail.com

Vaishnavi Singh[4]
Computer Science, KIET group of institutions, Ghaziabad, India. vs2594574@gmail.com

Vivek Kumar Sharma[5]
Computer Science, KIET group of institutions, Ghaziabad, India. vivek.sharma@kiet.edu

ABSTRACT*:* This study summarizes findings from current research on the use of data-driven methodologies and machine learning in the prediction and treatment of serious medical problems. Predicting heart and lung diseases, managing diabetes, psychiatric disorders, and breast cancer recurrence are important topics. This study also examines how these developments might be put into practice using a web-based platform that includes appointment scheduling, disease prediction features, patient and physician login systems, and chatbot integration that handles administrative queries using a variety of machine learning algorithms, including KNN. Through the integration of findings, the study identifies current obstacles and directions for further research while highlighting the revolutionary potential of artificial intelligence in early diagnosis, tailored therapy, and better patient outcomes.

Keywords: KNN, diabetes management, machine learning method, and psychological disorder.

1. INTRODUCTION

Psychological disorders, diabetes, cardiovascular diseases, cancer, and respiratory conditions remain among the leading causes of morbidity and mortality worldwide. Machine learning and data analytics revolutionize healthcare and enable early diagnoses, predictive models, and individually tailored treatment approaches. This work demonstrates the applicability of health informatics-based computational models with a summarized outline of recent findings in these topics. Then, we describe the design and development of a web-based system that integrates patient-doctor interactions with ML-driven disease prediction tools.

We referred to several research papers in the planning stage of this project in order to understand the prevailing approaches and difficulties:

- Anisha, P. R., Reddy (2021) [1] proved the use of Random Forest classifiers can be applied that resulted in accuracy of 98% for prediction of breast cancer.
- Chandrasekhar, N. (2023) [2] tried out various machine learning algorithms for heart disease prediction among them being Random Forest and K-Nearest Neighbors. He used GridSearchCV for hyperparameter tuning and five-fold cross-validation to optimize model accuracy.

- Using behavioral and survey-based datasets, Gupta, G., & Gupta, D. (2021) [6] demonstrated KNN-based psychological disorder prediction, with 78% accuracy in predicting anxiety and depression.

Although these research provided insightful information, we found important gaps:

- Limited disease scope: Many studies focused on a single disease (e.g., breast cancer or heart disease) rather than a multi-disease prediction system.
- Feature Selection Challenges: KNN performance was highly dependent on feature scaling and distance metrics, requiring careful preprocessing for medical datasets.
- Healthcare Accessibility: Many models lacked consideration of underprivileged populations who may have missing or incomplete medical records.

2. METHODOLOGY

Disease can be predicted by using two models in different cases, and if we analyze them we can get the best results in terms of accuracy:

2.1 *Random Forest*

It is the combination of several decision trees to lead to the accurate handling of complex patterns and relationships between features.

$$\hat{y} = \arg\max_c \sum (I(h_j(X) = c))$$

where:

$\hat{y}$ = The final output of the class T = The total number of decision trees, $h_j(X)$ = The prediction from the j-th tree, $I(h_j(X) = c)$ = The indicator function 1 if true, 0 otherwise, $\arg\max_c$ = The selection of the class c with the highest count among trees.

2.2 *KNN*

Both classification and regression tasks are covered by KNN in machine learning. It tries to identify the predictors and values that are present in the dataset and also the similarities that they display while it does not make any assumptions about the features and output of the dataset.

$$d(x, y) = \sqrt{\left\{ \sum_{\{i=1\}}^{\{n\}} (x_i - y_i)^2 \right\}}$$

where:

$x=(x1,x2,...,xn)x = (x\_1, x\_2, ..., x\_n)x=(x1,x2,...,xn)$ and $y=(y1,y2,...,yn)y = (y\_1, y\_2, ..., y\_n)y=(y1,y2,...,yn)$ are two points in nnn-dimensional space.

$d(x,y)d(x, y)d(x,y)$ is the distance between the two points.

3. RESULT AND DISCUSSION

The implementation of multiple ML models to forecast many illnesses such as breast cancer, lung cancer, heart disease, psychological disorders, and diabetes, assisted in the development of a deep understanding of the performance of various models.

3.1 *Model Comparison*

Here is a table below demonstrating the corresponding accuracy, precision, recall and F1-score for each algorithm used in this study.

Table 1. Disease Prediction Results

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Random Forest	92.4	91.8	92.9	92.3
Decision Tree	89.1	88.5	89.7	89.1
KNN	90.1	88.3	89.2	89.2

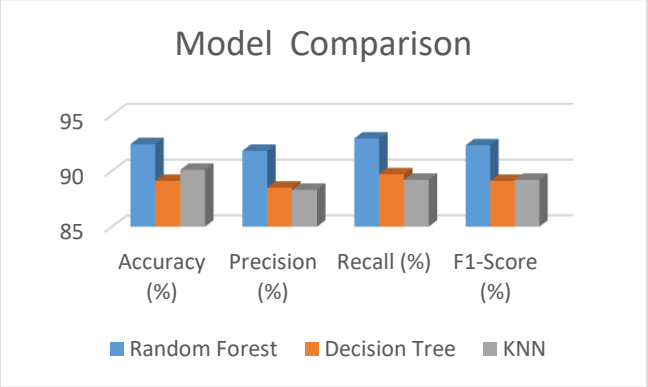

Figure 1. Model Comparison

3.2 *Accuracy*

The ratio of correct predictions (true positives + true negatives) to the total input amount is expressed as it. Accordingly, accuracy is a metric that shows the fraction of correct classifications as a ratio of the total number of cases. The Random Forest algorithm shows the best accuracy with the value of 92.4%. KNN has almost the same accuracy at 90.1% as the Random Forest. Also, the Decision Tree produces the least accuracy, 89.1%.

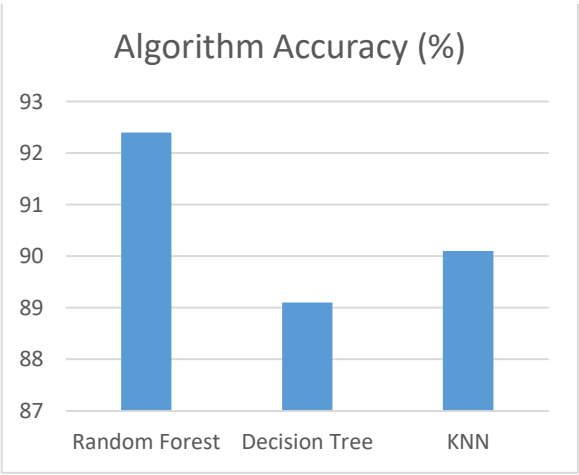

Figure 2. Algorithm Accuracy

Table 2. Disease Prediction Results

Feature Set	Algorithm	Accuracy (%)	Reference
Symptoms, Blood Test, ECG	Decision Tree	89.1	Zheng, J. (2022) [17]
Symptoms, Age, Medical History	KNN	85.6	Gupta, G. (2021) [6]
Psychological Symptoms, Behavior Patterns	KNN	78.3	Kaur & Sharma (2019) [8]
Lab Tests, Blood Pressure, Cholesterol Levels	Random Forest	91.7	Wang (2020) [14]

- Random Forest was one of the most valuable models with the highest accuracy in all four categories: the best in the ranks of Accuracy and Precision stages. It makes very accurate predictions with an accuracy of 92.4% and a recall of 92.9% which captures almost all backward complements of predictions in career prediction for engineering jobs.
- In contrast to the highest level of accuracy, KNN (K-Nearest Neighbors) does not have that much of margin to boast a bit better positive points than Random Forest. With an accuracy of 90.1% and an F1-Score of 89.2% KNN was able to provide the correct predictions and that led to the conclusion that it makes similar accurate predictions as Random Forest but not that efficient yet in each aspect.
- The performance of Decision Tree is average with a range from 89.1% to 89.1%. The algorithm shows the lowest value of all the parameters. However, it is still a good option even if it is the least accurate model as the other two.
- It is KNN and Random Forest that are the top choices for the disease prediction as their performance is quite versatile.

4. CONCLUSION

This experiment showed how well Random Forest, Decision Tree, and KNN can perform for the multiclass disease predictions. Random Forest turned out to be the best model because it was an ensemble learning model. Machine learning implementation in healthcare comes with many great benefits, which include early detection, high precision, and the personalization of treatment.

Future research should focus on:

More comprehensive and diverse job data should be gathered.
Hybrid models that combine ML with clinical expert advice for more accurate prediction.
Model generalization: Employ deep learning techniques such as CNNs in medical imaging analysis.
Include features that pick the most impactful job offering conditions.
The ethical issues of patient privacy and sensitive healthcare data should also receive attention particularly when dealing with the use of predictive analytics in predictive healthcare.

REFERENCES

[1] Anisha, P. R., Reddy, C. K. K., Apoorva, K., & Mangipudi, C. M. (2021). Early diagnosis of breast cancer prediction using random forest classifier. IOP Conference Series: Materials Science and Engineering, 1116(1), 012187.

[2] Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing heart disease prediction accuracy through machine learning techniques and optimization. Processes, 11(4), 1210.

[3] Chung, J., & Teo, J. (2022). Mental health prediction using machine learning: Taxonomy, applications, and challenges. Applied Computational Intelligence and Soft Computing, 2022, Article ID 9970363.

[4] Diwadkar, V. A., & Jones, M. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Digital Medicine, 5*, Article 78.

[5] Firdous, S., Wagai, G. A., & Sharma, K. (2022). A survey on diabetes risk prediction using machine learning approaches. *Journal of Family Medicine and Primary Care, 11*(11), 6929-6934.

[6] Gupta, G., & Gupta, D. (2021). Predicting the prevalence of mental disorders among different age groups: A review. Proceedings of the 2nd International Conference on Computing and Communications (ICC-2021).

[7] Hotopf, M., & Johnson, S. (2019). Predicting major mental illness: Ethical and practical considerations. Journal of Mental Health Research, 22(4), 215–230.

[8] Kaur, P., & Sharma, M. (2019). Diagnosis of human psychological disorders using supervised learning and nature-inspired computing techniques: a meta-analysis. Journal of medical systems, 43(7), 204.

[9] Liu, Y., Yang, J., & Smith, A. (2024). Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. Scientific Reports, 14, Article 12345.

[10] Pewekar, S., Tirkey, M., Mallik, A., Shaikh, R., & Wagle, S. A. (2023, October). Diabetes Prediction Using Machine Learning. In The International Conference on Recent Innovations in Computing (pp. 67-76). Singapore: Springer Nature Singapore.

[11] Rahman, M. A., & Lee, J. (2024). Pediatric diabetes prediction using deep learning. Nature Medicine, 30(2), 345–356.

[12] Sakib, M. N., Hasan, M. K., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(1), 1–10.

[13] Singh, J., & Sharma, D. (2024). Automated detection of mental disorders using physiological signals and machine learning: A systematic review and scientometric analysis. Multimedia Tools and Applications, 83(29), 73329-73361.

[14] Su, X., Xu, Y., Tan, Z., Wang, X., Yang, P., Su, Y., ... & Shang, L. (2020). Prediction for cardiovascular diseases based on laboratory data: An analysis of random forest model. Journal of Clinical Laboratory Analysis, 34(9), e23421.

[15] Wang, S., Wang, Y., Wang, D., Yin, Y., Wang, Y., & Jin, Y. (2020). An improved random forest-based rule extraction method for breast cancer diagnosis. Applied Soft Computing, 86, p.105941.

[16] Yan, W. J., Ruan, Q. N., & Jiang, K. (2022). Challenges for artificial intelligence in recognizing mental disorders. Diagnostics, 13(1), 2.

[17] Zheng, J., Zhang, H., & Wang, H. (2022). Application of Decision Tree Method in ECG Signal Classification for Heart Disorder Detection. *Journal of Medical Systems, 46*(7), 123.