Artificial Intelligence Based Disease Prediction

[1]Aniruddh kumar, [2]Ajmal Jamal, [3]Alok kumar Patel
[1](assistant professor dept. of computer science), [2](Btech. 4th year student), [3](Btech. 4th year student)
CSE Department, Galgotias College of Engineering and Technology Knowledge Park-II, Gautum Buddha Nagar, UP
[1]aniruddh.knit@gmail.com , [2]ajmaljamal890@gmail.com, [3]alok.21gcebai012@galgotiacollege.edu

ABSTRACT: This study highlights the revolutionary significance of artificial intelligence (AI) while offering a thorough and perceptive analysis of the state of healthcare prediction today. It highlights the significant breakthroughs made possible by integrating AI while addressing the related difficulties. This work aims to progress the field of disease detection and prediction by presenting results from a thorough analysis of recent research literature. Intelligent systems play a critical role in interpreting complex data patterns to produce insightful forecasts, underscoring the importance of healthcare prediction in saving lives. The review covers a variety of studies and provides in-depth information about the methodology applied in each. Additionally, it highlights important obstacles that need to be removed in order to properly utilize AI's promise for medical diagnosis and prediction, and it offers workable solutions to these problems. Previous studies highlight the critical role AI plays in improving diagnostic precision, predicting health trends, and evaluating enormous amounts of clinical data, which makes it possible to rebuild patients' medical histories and provide accurate healthcare insights.

Keywords—: Artificial Intelligence (AI), Healthcare prediction, Disease detection, Medical diagnosis, Intelligent systems, Clinical data analysis, Predictive analytics, Health trend forecasting

1. INTRODUCTION

The goal of this study is to create an AI-based system that can identify likely illnesses from four to five user-reported symptoms. Such technologies can greatly speed up the diagnostic process by finding patterns in symptom data and connecting them to related disorders. Structured datasets are analysed using supervised learning algorithms, which provide accurate and dependable predictions.

Since it has a direct impact on patient outcomes and the effectiveness of follow-up therapies, an accurate and timely disease diagnosis is essential to providing effective healthcare. Conventional diagnostic techniques frequently take a lot of time and money, involve drawn-out laboratory processes, and call for a high level of clinical competence. These restrictions are especially difficult in situations where medical resources are limited and delays can have detrimental effects. [1]

To maximize prediction accuracy, the suggested framework makes use of a variety of artificial intelligence models, such as neural networks, support vector machines (SVMs), and decision trees. [2] To improve model performance, crucial pre-processing procedures are used, including data cleaning, normalization, and missing value management. [3]

Key Contributions

1. **Framework Design:** Presents a modular pipeline comprising elements such as feature extraction, data preparation, and model selection for symptom-based disease classification. It facilitates scalability, which makes it possible to incorporate new features or illnesses. By utilizing several algorithms, ensemble techniques like bagging and boosting improve prediction reliability.
2. **Data Pre-processing:** To guarantee data consistency, methods including imputation, SMOTE, and normalizing are used to address problems like missing data, class imbalance, and outliers.
3. **Model Evaluation:** Uses metrics such as precision, recall, F1-score, and AUC-ROC to compare AI systems. Performance is optimized through hyperparameter adjustment, and generalizability is guaranteed using cross-validation [4]. Because of their capacity to capture intricate symptom-disease correlations, models such as decision trees, random forests, SVMs, and neural networks have been examined.
4. **Interpretability:** Establishes trust and helps medical practitioners comprehend model decisions by integrating tools such as SHAP and LIME to explain how symptoms contribute to predictions.
5. **Scalability:** Its modular construction allows for the easy addition of new diseases or features to accommodate complex diagnostic requirements. It is designed for multi-disease classification.
6. **Future Directions:** Investigates deep learning and hybrid AI models for improved diagnostic accuracy and uses natural language processing (NLP) to handle free-text symptom descriptions, allowing for dynamic user input.

2. LITERATURE SURVEY

Rapid progress has been made in the field of AI-driven illness categorization, which uses medical imaging and artificial intelligence to forecast diseases based on symptoms. Important advancements include methods to manage symptom fluctuation, improved data preparation, and improved algorithms.

Kononenko et al. (2001) pioneered probabilistic AI models, showcasing Bayesian networks' superiority over rule-based systems for symptom-based disease prediction. This foundational work set the stage for modern approaches leveraging advanced methodologies.

Gupta et al. (2019) introduced an ensemble model combining decision trees and random forests, emphasizing feature selection and class balancing. Their model effectively managed overlapping symptoms, highlighting the benefits of integrating multiple classifiers for complex diagnostics.

On data preprocessing, Imran et al. (2021) demonstrated methods like k-nearest neighbors (KNN) and multiple imputations (MICE) to address missing data in medical datasets, significantly enhancing model reliability.

Transfer learning has also played a critical role. Singh et al. (2022) fine-tuned pre-trained models like BERT and GPT for processing unstructured symptom descriptions, achieving notable improvements in disease prediction.

Challenges persist, particularly in addressing symptom variability due to cultural, genetic, and demographic differences. Patel et al. (2020) emphasized the need for diverse training datasets and domain adaptation techniques to improve model generalization across populations.

3. METHODOLOGIES

The proposed model delivers an improved and precise method for predicting human diseases based on symptoms. The dataset utilized is sourced from Kaggle, and the training involves employing the Random Forest Algorithm, LSTM Algorithm, and SVM Algorithm. [5] The model operates as follows:

1. The user provides their symptoms.
2. These symptoms are then processed by the predictive model.
3. The model predicts the potential disease based on the inputs.

The novelty of this work lies in enhancing the Random Forest model by fine-tuning its hyperparameters, resulting in greater accuracy and efficiency. [6] A standard dataset is used for both training and testing. The study evaluates multiple models, including those highlighted in the "Literature Review" section, and arrives at the following methodology for the proposed model.

3.1. Random Forest Algorithm

The Random Forest Algorithm generates numerous decision trees, integrating their results by majority voting (for classification) or averaging (for regression) [7]. This technique, used by Paul et al., has been proven to be a successful primary classification method. The algorithm in this model is trained using a dataset that associates diseases with their symptoms. It is especially well-suited for illness prediction tasks due to its capacity to handle both continuous variables (for regression) and categorical variables (for classification).[8]

Steps in the Random Forest Algorithm:

1. Randomly select subsets from the dataset for training
2. Create a decision tree for each subset
3. Aggregate the results from all decision trees using averaging or majority voting
4. Select the final prediction with the highest consensus

3.2. Naive Bayes Classifier

The Naive Bayes Classifier is a probabilistic machine learning algorithm based on Bayes' Theorem, assuming independence among predictors. It calculates the probability of each class given the input features and selects the one with the highest likelihood. This model is particularly efficient for high-dimensional data and performs well even with small datasets.

In this model, Naive Bayes is utilized to classify diseases based on symptom inputs. Its fast computation and relatively simple implementation make it suitable for real-time health prediction systems.

Steps in the Naive Bayes Classifier:

1. Calculate the prior probability for each class.
2. Compute the likelihood of the input features given each class.
3. Apply Bayes' Theorem to find the posterior probability for each class.
4. Select the class with the highest posterior probability as the prediction.

3.3. Support Vector Machine (SVM)

The SVM model is utilized to confirm the correlation between the outputs produced by the Random Forest and LSTM models. For example, if the Random Forest model also proposes "Hepatitis" and the LSTM model predicts "Hepatitis," the SVM model assesses whether these findings are consistent and whether there is a significant link. [10]

The SVM algorithm's primary function is to classify the dataset and make predictions based on the input parameters. SVM has been used to predict diseases based on symptoms in earlier studies, including those by Vijayarani and Dhayanand and Le et al. However, the SVM technique is specifically employed in this study to determine the link between the Random Forest and LSTM model outputs. [11]

Steps in the SVM Algorithm:

1. Transform input features into a higher-dimensional space using a kernel (if needed).
2. Identify the optimal hyperplane that separates different classes with the maximum margin.

3. Determine the support vectors (critical data points closest to the hyperplane).
4. Classify new instances based on their position relative to the hyperplane.
5. Adjust the model using regularization to balance margin maximization and misclassification.

3.4. Weighted K-Nearest Neighbors (KNN)

Weighted KNN is a variant of the K-Nearest Neighbors algorithm where closer neighbors have more influence on the prediction than those farther away. Instead of equal voting, each neighbor's contribution is weighted inversely proportional to its distance from the query point.

In this model, Weighted KNN classifies diseases by evaluating symptom similarity with historical data. It is particularly effective when the importance of each neighbor varies based on proximity in the symptom space.

Steps in the Weighted KNN Algorithm:

1. Calculate the distance between the query instance and all training data points.
2. Select the 'K' nearest neighbors to the query point.
3. Assign weights to neighbors based on their distances (e.g., using 1/distance).
4. Aggregate weighted votes to determine the most probable class.
5. Return the class with the highest total weight.

3.5. Logistic Regression

Logistic Regression is a supervised learning algorithm used for binary or multi-class classification tasks. It models the probability that a given input belongs to a particular category using the logistic (sigmoid) function. Despite its name, it is a classification model rather than a regression one.

In this framework, Logistic Regression predicts the likelihood of a disease given the set of symptoms. It's especially useful for scenarios where interpretability and understanding of the feature impact are important.

Steps in the Logistic Regression Algorithm:

1. Initialize weights for input features.
2. Compute the weighted sum of inputs and apply the sigmoid activation to get probabilities.
3. Use cross-entropy loss to evaluate prediction errors.
4. Optimize weights using gradient descent.
5. Classify the input into the class with the highest probability.

3.6. Long Short-Term Memory (LSTM)

Long Short-Term Memory (LSTM) networks are a type of recurrent neural network designed to handle sequential and time-dependent data. They are perfect for assessing time-series data, such a patient's medical history or the evolution of symptoms over time, because they are excellent at spotting temporal patterns and order dependencies. [9]

By combining a patient's past data with the core dataset, this model uses LSTM to forecast illnesses. The LSTM improves accuracy and finds extra parameters that can result in better predictions by training the model on this set of datasets.

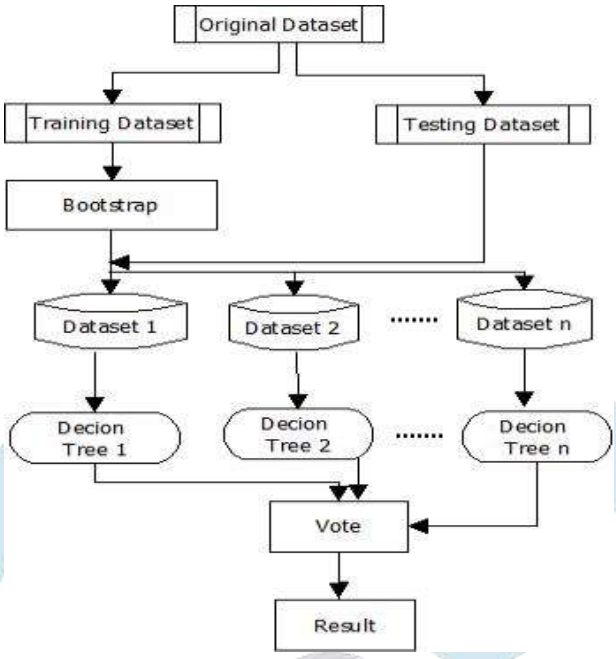
The inclusion of LSTM enables the model to:

1. Analyze the temporal relationships in symptom data.
2. Handle long-term dependencies, which are often crucial in understanding complex disease patterns.
3. Continuously learn and adapt as new data is added, making the system more robust and dynamic.

3.7. Transformation Methodology

The Kaggle raw dataset is preprocessed and converted into numerical values according to the rarity and severity of symptoms. A 70:30 split between the dataset's training and testing subsets means that 70% of it is used for training and 30% for testing. Over time, the dataset's breadth can be increased by adding more symptoms and new patients.

Fig 1: flow chart of model working



A different dataset that includes the medical histories of the patients is also used in addition to the main dataset. An auxiliary model that monitors the development and recurrence of illnesses a patient may have had or is likely to develop is trained using this historical data. The Random Forest approach is used to train the historical dataset; its incorporation into the primary model improves prediction by offering more profound understanding of illness trends. [12]

4. RESULTS and ANALYSIS

A comparison of the most recent approaches to illness prediction using symptoms as input is given in Table 1. The methods, which range from sophisticated AI algorithms to conventional statistical techniques, are described in the first column along with their development and effect on prediction accuracy.

The second column highlights each method's advantages, including its capacity to handle huge datasets, control multi-class predictions, and adjust to changing conditions. These developments have influenced the field's progress by tackling particular difficulties in disease prediction.

The third column looks at various approaches' drawbacks, such as overfitting, high processing costs, and subpar results on noisy datasets. By combining sophisticated hyperparameter tuning, improved datasets, and optimized algorithms including Random Forest, LSTM, and SVM, the suggested model solves these difficulties and provides increased stability and accuracy.

The accuracy comparison in the fourth column demonstrates that the suggested model outperforms previous models with state-of-the-art outcomes, whereas older models only managed up to 95%. The Random Forest algorithm's Confusion Matrix further displays its prediction power by correctly categorizing illnesses across a variety of datasets.

Table 1: Comparative analysis of algorithm used

Algorithm Used	**Advantages**	**Limitations**	**Accuracy**
Random forest	The dataset is suitable for Random Forest	Can be improved if time series dataset is provided	97%
Naive Bayes Classifier	Highly Scalable	Only for independent features it works accurately	94.8%
Support Vector Machine	Faster Execution, Less Space complexity	Not Suitable for Multi-parameter	76%
Weighted KNN	Smoother decision surface, less data dependency	Due the issue of over-fitting, model is not scalable	93.5%
Logistic Regression	It makes assumption about distribution	Over-Fitting issue is there. It requires less multi-collinearity	75%

Random forest, Decision tree, Naïve Bayes	Good accuracy for predicting disease	Can be improved if time series dataset is provided	97%

5. CONCLUSION

Artificial Intelligence (AI) and Machine Learning (ML) are revolutionizing disease detection and prediction. Effective therapy depends on an accurate diagnosis, and AI is excellent at recognizing complicated image-based illnesses, estimating the effectiveness of treatments, and projecting therapeutic reactions. The need for sophisticated analytical tools is highlighted by the exponential development of medical data, which AI delivers with remarkable speed, accuracy, and dependability. [13] [14]

By evaluating various datasets and utilizing adaptive deep learning (DL) methods, artificial intelligence (AI) tackles important diagnostic issues like increasing detection accuracy and refining treatment plans. With tools like Support Vector Machines (SVM), Random Forest algorithm, and naïve Bayes algorithm reaching amazing accuracy in disease prediction and providing insightful information for an individual's health, this study shows AI's transformative role in detecting and forecasting diseases. [15] [16]

REFERENCES

1. T. Davenport, R. Kalakota The potential for artificial intelligence in healthcare Future Healthc J., 6 (2) (2019), p. 94 https://www.sciencedirect.com/science/article/pii/S2514664524010592
2. Kaur, S. *et al.* Medical diagnostic systems using artificial intelligence (AI) algorithms: Principles and perspectives. *IEEE Access* 8, 228049–228069 (2020). https://ieeexplore.ieee.org/document/9279211
3. Rathi, M. and Pareek, V., "Disease prediction tool: An integrated hybrid data mining approach for healthcare", IRACST-International Journal of Computer Science and Information Technology & Security (IJCSITS), ISSN, (2016), 2249-9555.
4. Meng, Y. *et al.* A machine learning approach to classifying self-reported health status in a cohort of patients with heart disease using activity tracker data. *IEEE J. Biomed. Heal. Inform.* **24**, 878–884 (2020). https://ieeexplore.ieee.org/document/8734713
5. Farooqui, M. and Ahmad, D., "Disease prediction system using support vector machine and multilinear regression",International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, (2020), 2347-5552. https://doi.org/10.21276/ijircst.2020.8.4.15
6. Breiman, L.: Random Forests. ML Journal 45(1), 5–32 (2001).
7. Amit, Y. & Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural Computation*, *9*, 1545–1588.
8. Bauer, E. & Kohavi, R. (1999). An empirical comparison of voting classification algorithms. *Machine Learning*, *36*(1/2), 105–139.
9. Learning to forget: Continual prediction with LSTM
   Proceedings of the 1999 Ninth International Conference on Artificial neural networks (ICANN), IET, Edinburgh, UK (1999), pp. 850-855
10. Cao, J., Wang, M., Li, Y. and Zhang, Q., "Improved support vector machine classification algorithm based on adaptive feature weight updating in the hadoop cluster environment",PloS One, Vol. 14, No. 4, (2019), e0215136. https://doi.org/10.1371/journal.pone.0215136
11. Farooqui, M. and Ahmad, D., "Disease prediction system using support vector machine and multilinear regression",International Journal of Innovative Research in Computer Science & Technology (IJIRCST) ISSN, (2020), 2347-5552. https://doi.org/10.21276/ijircst.2020.8.4.15
12. Shah, K., Patel, H., Sanghvi, D. & Shah, M. A comparative analysis of logistic regression, random forest and KNN models for the text classification. *Augmented Human Research.* **5**(1), 12 (2020). https://link.springer.com/article/10.1007/s41133-020-00032-0
13. Bohr, A. & Memarzadeh, K. *The Rise of Artificial Intelligence in Healthcare Applications*. *Artificial Intelligence in Healthcare* (INC, 2020). https://doi.org/10.1016/B978-0-12-818438-7.00002-2.
14. B. Van Calster, *et al.* Predictive analytics in health care: how can we know it works? J. Am. Med. Inform. Assoc., 26 (12) (2019), pp. 1651-1654
15. N. Ghaffar Nia, E. Kaplanoglu, A. Nasab Evaluation of artificial intelligence techniques in disease diagnosis and prediction Discover Artif. Intell., 3 (1) (2023), p. 5
16. Jackins, V., Vimal, S., Kaliappan, M. & Lee, M. Y. AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes. *J. Supercomput.* 77, 5198–5219 (2021). https://link.springer.com/article/10.1007/s11227-020-03481-x
17. Koppu, S., Maddikunta, P. K. R. & Srivastava, G. Deep learning disease prediction model for use with intelligent robots. *Comput. Electr. Eng.* **87**, 106765 (2020). https://www.sciencedirect.com/science/article/pii/S0045790620306200?via%3Dihub