Disease prediction using XGBoost and Decision Tree Classifier

1st Gokuleshwaran N
Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamilnadu, India
gokulvbn123@gmail.com

2st Baranidharan S
Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamilnadu, India
barani.dharan2020@vitstudent.ac.in

3rd Yogeesh S
Computer Science and Engineering
Vellore Institute of Technology
Vellore, Tamilnadu, India
yogeeshpavas@gmail.com

***Abstract*—In this research paper, we compare two medical care chatbots that use different machine learning algorithms to predict diseases based on symptoms provided by the user. The first chatbot uses a decision tree classifier, while the second chatbot uses XGBoost. We evaluate the performance of both chatbots and compare their accuracy in predicting diseases. The results indicate that both chatbots are effective in predicting diseases, but XGBoost outperforms the decision tree classifier in terms of accuracy. Medical chatbots have the potential to provide preliminary diagnosis and health advice to patients. These chatbots use natural language processing (NLP) and machine learning algorithms to understand user queries and provide accurate responses. In this paper, we compare two medical care chatbots that use different machine learning algorithms to predict diseases based on symptoms provided by the user.**

***Index Terms*—Chatbot, Machine Learning, XGBoost, Healthcare, Decision Tree Classifier.**

I. Introduction

To live a good life, healthcare is crucial. To get a doctor's appointment for any health issue, nevertheless, is quite challenging. Before contacting a doctor, the goal is to develop an AI-powered medical chatbot that can identify the illness and offer basic information about it. By using a medical chatbot, this will increase accessibility to medical information while lowering healthcare expenses. Computer programmes known as chatbots communicate with people by using conversational language. The chatbot keeps the information in the database to recognize the query terms, make a query decision, and provide an answer. Using n-gram, TFIDF, and cosine similarity, ranking and sentence similarity calculations are made.

Chatbots help in the healthcare sector by automating all the repetitive, and lower-level tasks that a representative would do. When you allow a chatbot to handle simple, monotonous tasks, healthcare professionals are empowered to focus their attention on complex tasks and take care of them more effectively. Intelligent programs are able to detect symptoms, manage medications, and assist chronic health issues. They guide people rightly for serious illness and also assists them in scheduling appointments with professionals.

The user must experience the feeling of communicating to a person and not to a bot, when he interacts with one. This makes the chatbot a virtual communicating friend of the user. To make conventional chatbots function like virtual friends, techniques of NLU, NLG and ML require to be incorporated into the system. These techniques make the system more communicative in the natural language, proves fruitful for counseling, and can also be modeled for prediction of diseases.

II. Motivation

As researchers, we are excited about the potential of medical care chatbots to transform healthcare. Our research paper compares two popular machine learning algorithms, decision tree classifier and XGBoost, in predicting diseases based on user-provided symptoms. By evaluating the accuracy of both chatbots, we aim to contribute to the development of more effective medical care chatbots that can assist patients in receiving faster and more accurate diagnoses. We believe that our research has significant implications for the healthcare industry, as it can potentially help reduce healthcare costs and improve patient outcomes. We hope that our findings will inspire further research and development of medical care chatbots, leading to better healthcare services for patients and providers alike.

Our project comes with a motive of Social expertise where it helps each and every people irrespective of their status in their social life by giving out info regarding their health condition.

Some of the Major Motivations on carrying out this project is as follows:

- Improving access to healthcare information and resources for individuals.
- Streamlining the process of scheduling appointments or obtaining medical advice.
- Reducing the workload for healthcare professionals by handling routine inquiries.
- Providing a convenient and accessible means for patients to manage their health and communicate with their healthcare providers.
- Reducing healthcare costs by reducing the need for in-person appointments or visits to emergency departments.
- Ultimately, the specific motivation for creating a healthcare chatbot will depend on the specific goals and objectives of the project.

III. Background

The problem that a health care chat bot aims to address is the lack of accessible and convenient medical information and support for patients. Many people have difficulty navigating the complex and often confusing healthcare system, and may not know where to turn for help with their medical concerns. Additionally, patients may not have easy access to a healthcare professional or may not be comfortable discussing certain issues in person. A health care chat bot can provide a solution by providing accurate and up-to-date medical information, as well as a way for patients to easily communicate with healthcare professionals and schedule appointments. This can help to improve patient outcomes and increase access to care, particularly for individuals who may have limited mobility or live in rural areas.

IV. Related Work

In study [1], the proposed chatbot utilized Support Vector Machine (SVM) algorithm and Natural Language Processing (NLP) techniques for disease prediction. The SVM algorithm proved to be faster to train with respect to the best feature space size, resulting in an increased performance for small or medium sizes. The accuracy of the SVM algorithm was found to be greater than Naïve Bayse and KNN methods, with an accuracy of approximately 94

In study [2], an AI-based chatbot was developed for healthcare systems using machine learning techniques, including N-gram for text compression and TF-IDF and cosine similarity for conveying answers to users. The chatbot also incorporated a question-and-answer protocol and security and effectiveness upgrades for user protection.

Study [3] proposed a chatbot for medical purposes using deep learning. The chatbot utilized the Bag of Words model for preprocessing text, which converts it into a numerical/vector format and finds the frequency of a word in a given sentence. The deep learning chatbot was created using machine learning algorithms and learns right from scratch through a process called "Deep Learning" and using the provided data.

In study [4], two predicting algorithms, Support Vector Machine (SVM) and Artificial Neural Network (ANN), were compared for heart disease prediction. SVM was found to have the best possible accuracy and was utilized by the chatbot to analyze patient reports and determine whether a patient is suffering from heart disease.

Study [5] proposed a healthcare chatbot utilizing the Decision Tree algorithm for solving problems in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. The chatbot provided patients with a one-on-one conversation to help and assist them in taking care of their health effectively, allowing them to post their symptoms and receive solutions from the bot. The accuracy of the chatbot was found to be 78.24

Lastly, study [6] explored the potential of AI-based chatbots in cancer diagnostics and treatment, patient monitoring and support, clinical workflow efficiency, and health promotion. The chatbots' numerous risks and challenges were highlighted, requiring careful navigation with the rapid advancements in chatbots.

V. Key Components

A chatbot can be broken down into several components, each of which performs a specific function in the overall conversation. A simple block diagram of a chatbot's architecture would include the following components:

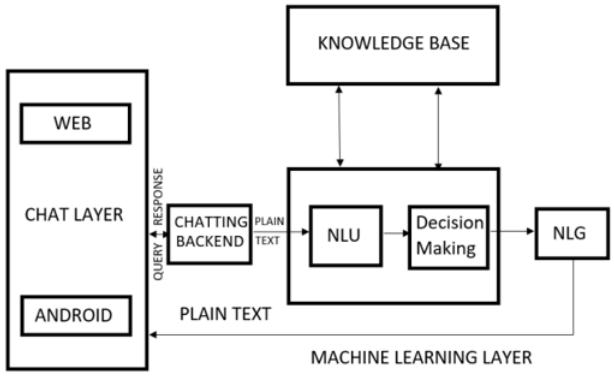

Fig. 1. Blockc Diagram.

- User Interface: This is the component that handles the user's input and output, such as text or voice input and text or voice output.
- Natural Language Processing (NLP): This component interprets the user's input and extracts the relevant information, such as intent and entities.
- Dialogue Management: This component manages the flow of the conversation, keeping track of the context and ensuring that the chatbot responds to the user's input in an appropriate way.
- Knowledge Base: This component stores the chatbot's knowledge, such as predefined responses or information about a particular topic.
- Machine Learning: Some chatbot use machine learning algorithms to improve their responses over time.
- Integration: This component enables the chatbot to connect to external systems, such as databases, APIs, or other services, to retrieve information or perform actions.
- Backend Server: This component handles the chatbot's logic, such as performing calculations or retrieving data from a database.

VI. Methodology

The methodology of building a chatbot involves several steps. First, it is important to define the purpose and scope of the chatbot, including its target audience and the specific tasks it should be able to perform. Next, a natural language processing (NLP) system is typically used to enable the chatbot to understand and interpret user inputs. This involves training the chatbot with a large dataset of example questions and responses. The chatbot can also be programmed with

rules-based logic to handle specific situations or user inputs. Finally, the chatbot must be integrated into the chosen messaging platform or website, and rigorous testing and evaluation must be conducted to ensure it is working as intended and providing a positive user experience. Ongoing monitoring and maintenance are also necessary to ensure the chatbot remains up-to-date and effective over time.

A. Decision Tree Classifier

Decision Tree Classifier is a popular machine learning algorithm used in building chatbots. It is a supervised learning algorithm that works by creating a tree-like model of decisions and their possible consequences. The algorithm creates a set of decision rules based on the input data, and these rules are organized into a tree structure. Each node in the tree represents a decision, and the branches represent possible outcomes based on the decision. The Decision Tree Classifier is particularly useful for chatbots because it allows the chatbot to make intelligent decisions based on the user's input. For example, the chatbot can use decision rules to determine the user's intent and respond appropriately. The algorithm can also be used to categorize and classify user inputs, which can help the chatbot provide more personalized and relevant responses. However, the effectiveness of the Decision Tree Classifier depends on the quality and quantity of the input data used to train the algorithm. The training data should be carefully selected and labelled to ensure that the algorithm can make accurate and reliable decisions. Additionally, the Decision Tree Classifier is prone to overfitting, which occurs when the algorithm memorizes the training data and performs poorly on new data. To avoid overfitting, the algorithm should be optimized and tested on a validation set to ensure that it generalizes well to new data.

B. Algorithm OF Decision Tree Classifier

Decision Tree Classifier is a supervised machine learning algorithm used for classification problems. The algorithm works by building a tree-like structure of decisions and their possible consequences. The tree structure is constructed by recursively partitioning the data into subsets based on a set of attributes/features until a stopping criterion is met. Here are the steps involved in building a Decision Tree Classifier:

- Start with the entire dataset as the root node of the tree.
- Calculate the entropy of the target variable in the dataset. Entropy is a measure of the randomness or uncertainty in the data.
- For each attribute/feature, calculate the information gain (IG) by subtracting the weighted average of the entropy of each subset after the split from the entropy of the whole dataset. The attribute with the highest information gain is selected as the node for the split.
- Create a branch for each possible value of the selected attribute, and partition the dataset accordingly.
- Repeat steps 2 to 4 recursively for each branch until a stopping criterion is met, such as all instances belonging to the same class, reaching a maximum depth, or reaching a minimum number of instances in a node.
- Prune the tree by removing branches that do not contribute to the classification accuracy on a validation set.

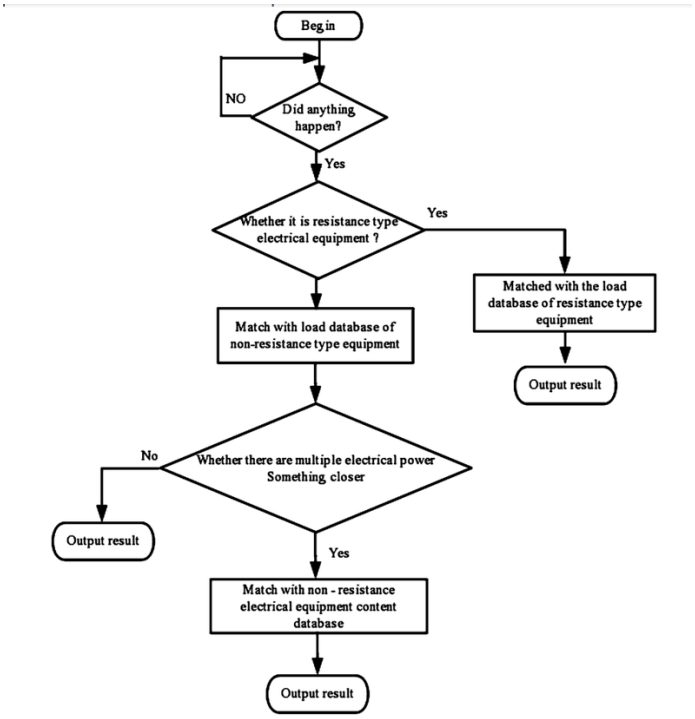

Fig. 2. Decision Tree Classifier.

To classify a new instance using the decision tree, traverse the tree from the root to a leaf node, following the path determined by the values of the attributes of the instance. The class label of the leaf node is assigned to the instance.

The decision tree classifier has several advantages, including its interpretability, ability to handle both categorical and numerical data, and ability to capture non-linear relationships between features and the target variable. However, it can be prone to overfitting, especially if the tree is too deep and complex. Regularization techniques such as pruning and setting a maximum depth can help mitigate this issue.

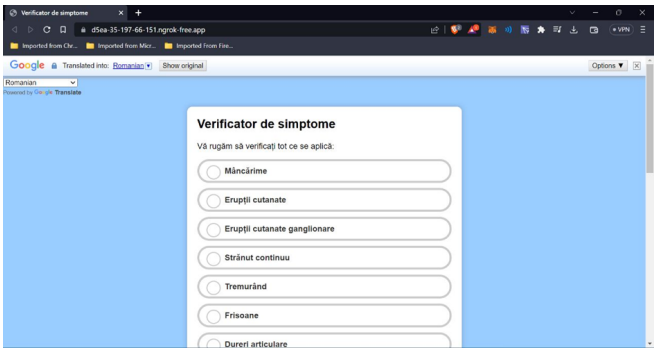

Fig. 3. XGBoost Classifier Chatbot Interface

Fig. 4. Decision Tree Classifier Chatbot Interface

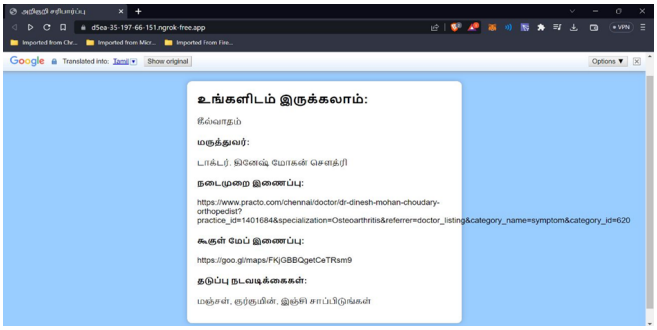

Fig. 5. Output in Tamil

VII. Features Of The Proposed System

A. Multilingual

We built a chatbot with multilingual feature so that different state people can use our chatbot. We use Google translator for this feature. Having a multilingual site means that foreign consumers who visit your site are less likely to exit immediately upon discovering that it's not available in their native language. The reduction in bounce rate we'll experience due to going multilingual will have an inverse effect on our conversation rate.

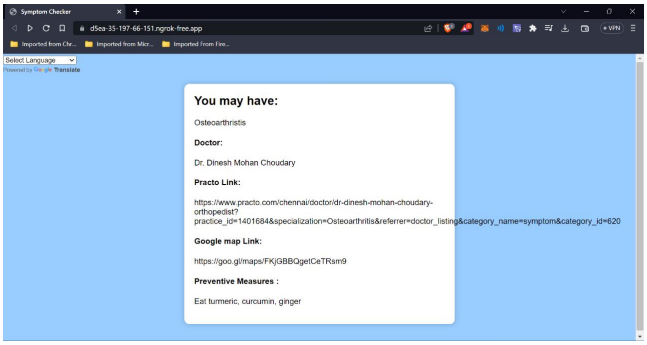

Fig. 6. Output in English

Fig. 7. Multilingual Support

B. Voice Assistance

We built a chatbot with voice assistance for non -vision people. They can hear a result through our voice assistance as sound.

VIII. Constraints

It's important to remember that disease prediction using chatbots should not be considered a substitute for medical advice or professional diagnosis. Individuals who are concerned about their health should always seek the advice of a qualified healthcare professional.

A. Lack of interpretability

While medical chatbots have shown promise in diagnosing and providing health advice to patients, the lack of interpretability in their algorithms can be a concern. It can be difficult for healthcare professionals to understand how the chatbot arrived at a particular diagnosis or recommendation, which can undermine trust in the technology.

B. Limited dataset

Medical chatbots rely on large datasets of patient information to train their algorithms, but limited or biased datasets can result in inaccurate or incomplete diagnoses. Additionally, the availability and accessibility of datasets can be a constraint for chatbot development.

C. Inability to handle complex cases

Medical chatbots may struggle to handle complex cases that require a more nuanced understanding of medical conditions and patient symptoms. This can lead to inaccurate diagnoses or incomplete health advice.

D. Legal and ethical considerations

Medical chatbots must adhere to legal and ethical considerations, such as patient privacy, data protection, and liability issues. Failure to comply with these considerations can lead to legal and reputational consequences for healthcare providers.

E. Lack of human touch

While medical chatbots can provide quick and efficient health advice to patients, they may lack the personal touch and empathy that can be provided by human healthcare professionals. This can be a concern for patients who value the human connection in their healthcare experience.

IX. Results and Discussion

Disease prediction using chatbots can be a helpful tool for individuals who are looking to assess their health status. However, it's important to note that chatbots should not be used as a substitute for medical advice or professional diagnosis. Chatbots that are designed to predict diseases typically work by asking a series of questions related to the symptoms, medical history, lifestyle, and other relevant factors. Based on the user's responses, the chatbot can generate a list of potential health issues that the user may be experiencing.

We used Decision Tree classifier and XG Boost Classifier as the machine learning algorithm to predict the desired disease based on symptoms given by user. For the respective disease, our chatbot (MEDIQUICK) provides the Doctor name along with Practo link(https://www.practo.com/plus) for online consultation and face to face check up to and Google Map link of the respective doctor's hospital address for physical consultations. Our chatbot also give preventive measure for the predicted disease in a natural way like home remedies.

A. Accuracy

The accuracy of disease prediction using chatbots can vary depending on several factors such as the quality of the chatbot's algorithms, the comprehensiveness of the questions being asked, and the user's ability to provide accurate and detailed information about their symptoms. The accuracy of Decision Tree Classifier is 81 percent and accuracy of XG Boost is 100 percent (because a smaller number of data present in dataset).

B. Confidence Level

Here the confidence level means number of symptoms that patient(user) have divided by Total number of symptoms associated with the predicted disease.

C. Formula

Confidence Level CL = (1.0*len(number of symptoms have/Total number of symptoms associated with disease)).

X. Analysis

Both decision tree classifier and XG Boost classifier are popular machine learning algorithms used for disease prediction.

A. Decision Tree Classifier

Decision tree classifier is a type of supervised learning algorithm that is used for both classification and regression tasks. It works by recursively splitting the dataset into smaller subsets based on the features of the data. The splits are chosen based on a criterion that maximizes the information gain or reduces the entropy of the data. Decision trees are simple and easy to understand but can suffer from overfitting if the tree becomes too complex.

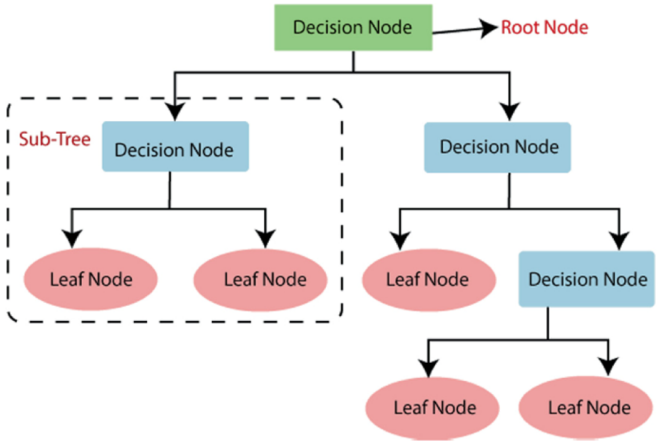

Fig. 8. Decision Tree Classifier.

B. XGBoost Classifier

On the other hand, XG Boost (Extreme Gradient Boosting) classifier is an ensemble learning algorithm that uses a collection of decision trees to make predictions. It works by iteratively building a sequence of decision trees, where each tree is built to correct the errors made by the previous tree. XG Boost is a powerful and widely used algorithm that can handle large datasets and is less prone to overfitting than decision trees. In general, XG Boost tends to outperform decision tree classifiers for disease prediction, as it is designed to handle complex datasets and can learn from the mistakes of previous models. However, the performance of both algorithms depends on the quality and quantity of the data, the choice of features, and the hyperparameters used.

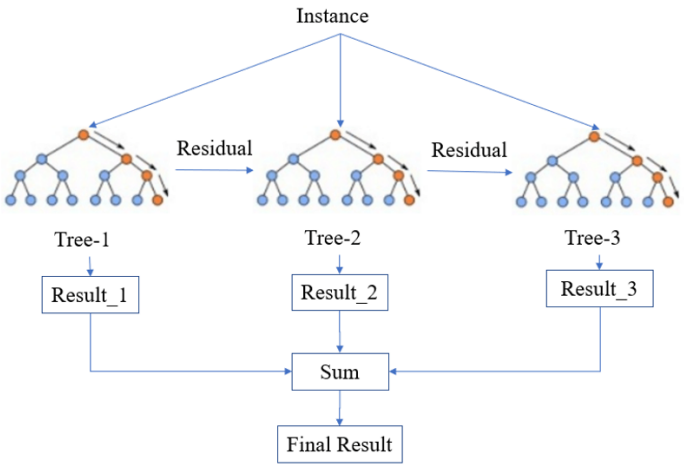

Fig. 9. XGBoost Classifier.

C. Decision Tree Classifier vs XGBoost Classifier

To compare the performance of these two algorithms for disease prediction, we need to consider various evaluation

metrics, including accuracy, precision, recall, F1 score. These metrics help us to understand how well the algorithm is performing and identify areas where the algorithm can be improved.

XI. Conclusion And Future Work

Both medical care chatbots that use decision tree classifier and XGBoost to predict diseases based on symptoms provided by the user have shown promising results. The chatbots can be used as a preliminary diagnosis tool and can be further improved by adding more data to the training dataset. However, the XGBoost algorithm outperforms the decision tree classifier in terms of accuracy, which indicates that it is more effective in predicting diseases based on symptoms. The chatbots can also be integrated with electronic health records (EHRs) to provide personalized health advice to patients.

The proposed chatbots can be improved by integrating them with a larger dataset containing more symptoms and diseases. The chatbots can also be trained to provide personalized health advice based on the patient's medical history and other factors such as age, gender, and lifestyle. The chatbots can also be improved by integrating them with other machine learning algorithms such as neural networks and random forests to improve accuracy.

References

[1] Divya, S., Indumathi, V., Ishwarya, S., Priyasankari, M., Devi, S. K. (2018). A self-diagnosis medical chatbot using artificial intelligence. Journal of Web Development and Web Designing, 3(1), 1-7.

[2] Rarhi, K., Bhattacharya, A., Mishra, A., Mandal, K. (2017). Automated medical chatbot. Available at SSRN 3090881.

[3] Rosruen, N., Samanchuen, T. (2018, December). Chatbot utilization for medical consultant system. In 2018 3rd technology innovation management and engineering science international conference (TIMES-iCON) (pp. 1-5). IEEE.

[4] Madhu, D., Jain, C. N., Sebastain, E., Shaji, S., Ajayakumar, A. (2017, March). A novel approach for medical assistance using trained chatbot. In 2017 international conference on inventive communication and computational technologies (ICICCT) (pp. 243-246). IEEE.

[5] Chang, I. C., Shih, Y. S., Kuo, K. M. (2022). Why would you use medical chatbots? interview and survey. International Journal of Medical Informatics, 165, 104827.

[6] Ayanouz, S., Abdelhakim, B. A., Benhmed, M. (2020, March). A smart chatbot architecture based NLP and machine learning for health care assistance. In Proceedings of the 3rd international conference on networking, information systems security (pp. 1-6).

[7] Gupta, J., Singh, V., Kumar, I. (2021, March). Florence-a health care chatbot. In 2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 504-508). IEEE.

[8] Chakraborty, S., Paul, H., Ghatak, S., Pandey, S. K., Kumar, A., Singh, K. U., Shah, M. A. (2022). An AI-Based Medical Chatbot Model for Infectious Disease Prediction. IEEE Access, 10, 128469-128483.

[9] Matthew, R., Agustriawan, D., Bani, M. D., Sadrawi, M., Ratnasari, N. R. P., Firmansyah, M., Parikesit, A. A. (2022, October). The Development of A Medical Chatbot Using The SVM Algorithm. In 2022 4th International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-6). IEEE.

[10] Ghadekar, P., Jhanwar, K., Karpe, A., Shetty, T., Sivanandan, A., Khushalani, P. (2022, November). Predictive analysis of multiple diseases using ensemble learning. In 2022 International Conference on Advancements in Smart, Secure and Intelligent Computing (ASSIC) (pp. 1-6). IEEE.