International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 10 | Oct - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Thyroid Disease Prediction Using Data-Driven Machine Learning Algorithms

M.Sravanthi[1],Muddasani Nanditha[2],Mandaloju SaiVardhan[3],Athram Gnaneshwar[4],Guglavath Swapna[5]

[1,2,3,4,5]Department of ECE, CMR Technical Campus, Hyderabad ,Telangana, India
Corresponding mail: nandithamuddasani27@gmail.com

ABSTRACT: Thyroid disorders, including hypothyroidism and hyperthyroidism, can be difficult to distinguish because symptoms may overlap and clinical interpretation requires multiple patient attributes and laboratory measurements. This paper presents a data-driven machine learning framework for classifying thyroid status into three categories: hypothyroid, hyperthyroid, and normal. The project workflow uses a thyroid dataset containing 3,772 records and 30 columns, followed by data preprocessing and an 80:20 train-test split (3,017 training records and 755 testing records). Decision Tree, Random Forest, Support Vector Machine (SVM), and Multilayer Perceptron (MLP) classifiers are evaluated and compared. The project report records approximately 99.20% accuracy for Decision Tree, 95% for SVM, and 98% for MLP; Random Forest is also reported at approximately 99% accuracy and is identified, together with Decision Tree, as one of the strongest models. The results indicate that tree-based classifiers are promising for this dataset. The system is intended as a decision-support aid and does not replace professional medical diagnosis.

Keywords - Thyroid disease prediction, machine learning, Random Forest, Decision Tree, SVM, MLP, classification, clinical decision support, data preprocessing.

I. INTRODUCTION

The thyroid gland regulates several important physiological processes through the production of thyroid hormones. Disorders such as hypothyroidism and hyperthyroidism can affect metabolism and general health, while symptoms may be subtle or overlap with other conditions. Diagnosis typically combines clinical assessment with laboratory measurements and professional interpretation. A data-driven classification system can help analyze patient attributes consistently and provide an additional source of decision support. Machine learning (ML) methods learn relationships between input features and known outcome labels. In thyroid prediction, relevant attributes may include hormone measurements such as TSH, T3, and T4, together with demographic or clinical variables available in the dataset. The project described here compares several classifiers and focuses on the three output classes hypothyroid, hyperthyroid, and normal. The objective is not to replace clinicians, but to support timely review of patient data and make model performance measurable.

II. LITERATURE SURVEY

This paper presents a machine learning approach for early prediction of hypothyroidism and multiclass classification of thyroid diseases. The study evaluates algorithms such as Decision Tree, Random Forest, Naive Bayes, and Artificial Neural Networks. Experimental results demonstrate that Decision Tree and Random Forest can achieve high classification accuracy, indicating their effectiveness in identifying thyroid disorders. The findings highlight the importance of selecting suitable classification algorithms for accurate thyroid disease prediction.[1].

This paper proposes a machine learning framework for thyroid disease classification using feature selection and risk factor identification. The study evaluates algorithms such as Support Vector Machine (SVM), AdaBoost, Decision Tree, Gradient Boosting, K-Nearest Neighbors, and Random Forest. Cross-validation is used to compare the performance of the models. The results demonstrate that feature selection and ensemble learning can improve classification performance and help identify important clinical attributes associated with thyroid disorders.[2].

This paper investigates thyroid disease classification using deep learning and traditional machine learning techniques. The study examines algorithms such as Decision Tree, Naive Bayes, Random Forest, and Support Vector Machine, along with neural-network-based approaches. The experimental findings demonstrate that combining feature selection with machine learning techniques can improve predictive performance. The study highlights the potential of neural networks for identifying complex patterns in clinical data and supporting thyroid disease classification.[3].

This paper presents an optimized machine learning approach for thyroid disease prediction using bio-inspired computing techniques. The study evaluates algorithms such as Random Forest, Decision Tree, Support Vector Machine, and K-Nearest Neighbors on thyroid datasets. A Particle Snake Swarm Optimization technique is incorporated to improve model performance. Experimental results demonstrate that the optimized Random Forest model achieves high classification accuracy across the evaluated datasets, highlighting the benefits of optimization techniques for improving thyroid disease prediction.[4].

This paper proposes a comprehensive framework for thyroid disorder diagnosis using feature selection, genetic algorithms, and machine learning techniques. The study investigates algorithms such as Logistic Regression, Random Forest, Support Vector Machine, AdaBoost, and Decision Tree. The framework focuses on identifying relevant clinical features and optimizing model performance. The results emphasize the importance of feature selection and performance evaluation using accuracy, precision, recall, and F1-score to develop effective thyroid disease prediction systems.[5].

This paper presents a comparative analysis of machine learning models for thyroid disease detection, focusing on prediction accuracy, interpretability, and clinical integration. The study evaluates algorithms such as Support Vector Machine, Decision Tree, Random Forest, Logistic Regression, and K-Nearest Neighbours. The findings indicate that Decision Tree and Random Forest can achieve strong classification performance on the evaluated dataset. The study also highlights the importance of thyroid-stimulating hormone (TSH) and other clinical attributes in predicting thyroid disorders.[6].

This paper proposes a stacking ensemble model for thyroid disease prediction by combining Random Forest, XGBoost, and

Decision Tree classifiers with a Logistic Regression meta-classifier. The framework incorporates feature selection and explainable artificial intelligence techniques such as SHAP and LIME to interpret the prediction results. Experimental results demonstrate high classification accuracy, while the explainability techniques provide insights into the factors influencing model predictions. The study highlights the potential of ensemble learning for improving prediction performance and interpretability.[7].

This paper presents a comparative empirical analysis of machine learning techniques for robust thyroid disorder detection. The study evaluates algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, XGBoost, LightGBM, and neural networks. The researchers investigate the influence of data preprocessing, class imbalance, and ensemble learning on classification performance. The findings demonstrate that ensemble methods can provide reliable predictive results and emphasize the importance of selecting appropriate evaluation metrics for thyroid disease classification.[8].

III. PROPOSED SYSTEM

The proposed system presents a data-driven machine learning framework for thyroid disease prediction by analysing patient-related clinical attributes and classifying individuals into hyperthyroid, hypothyroid, and normal categories. The system uses a dataset containing 3,772 records and applies data preprocessing techniques to handle numerical and non-numerical attributes, prepare the input features, and improve data consistency before model training. The pre processed dataset is divided into training and testing sets using an 80:20 split, with 3,017 records used for training and 755 records used for testing.

The proposed framework implements four machine learning algorithms: Decision Tree, Support Vector Machine (SVM), Random Forest, and Multilayer Perceptron (MLP). Each model is trained using the prepared training dataset and evaluated on the testing dataset to compare its predictive performance. Evaluation metrics such as accuracy, precision, recall, and F1-score are used to assess classification effectiveness. Confusion matrices are also generated to analyze correct predictions and misclassifications across the thyroid disease categories.

The Decision Tree algorithm provides an interpretable classification approach, while SVM identifies decision boundaries between different classes. Random Forest combines multiple decision trees to improve prediction robustness, and MLP learns complex relationships between clinical attributes through neural-network-based processing. A comparative evaluation is performed to identify the most suitable algorithm for the given dataset rather than relying on a single classification technique.

The system also provides a user-friendly interface through which users can upload the dataset, perform model training, visualize evaluation results, compare algorithm performance, and generate predictions for test records. Graphical representations of classification metrics and confusion matrices help users understand the behavior of the trained models. The modular design allows different algorithms to be evaluated within a common workflow. The proposed system aims to support early identification of potential thyroid disorders through automated classification of clinical data. By combining data preprocessing, multiple machine learning algorithms, performance evaluation, and prediction visualization, the framework provides a systematic approach to thyroid disease prediction. However, the predictions are intended to support research and preliminary screening and should not replace professional medical evaluation or laboratory-based diagnosis.

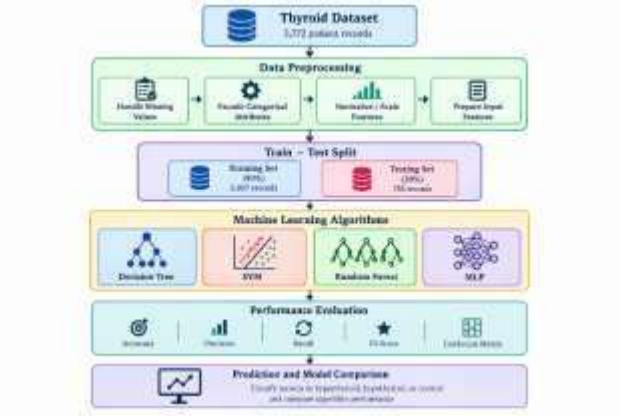


FIGURE 1: PROPOSED SYSTEM BLOCK DIAGRAM

The Figure 1 illustrates the overall workflow of the proposed thyroid disease prediction system. The system uses a dataset containing 3,772 patient records to train and evaluate data-driven machine learning algorithms. Data preprocessing techniques are applied to handle missing values, encode categorical attributes, normalize features where required, and prepare the input data. The pre processed dataset is divided into training and testing sets using an 80:20 split, consisting of 3,017 training records and 755 testing records. Four machine learning algorithms—Decision Tree, Support Vector Machine (SVM), Random Forest, and Multilayer Perceptron (MLP)—are used to classify thyroid conditions. The models are evaluated using accuracy, precision, recall, F1-score, and confusion matrices. Finally, the system compares model performance and predicts whether a record belongs to the hyperthyroid, hypothyroid, or normal category.

HARDWARE AND SOFTWARE IMPLEMENTATION

A. HARDWARE IMPLEMENTATION

The proposed thyroid disease prediction system is a software-based application that runs on a computer or laptop. The system uses the computer's processor, RAM, and storage to process the thyroid dataset, train machine learning models, and generate predictions. No dedicated hardware components or sensors are required.

B. SOFTWARE IMPLEMENTATION

The system is implemented using Python and machine learning libraries. Data preprocessing techniques are applied to handle missing values, encode categorical attributes, and prepare input features. The dataset is divided into training and testing sets using an 80:20 ratio. Decision Tree, SVM, Random Forest, and MLP algorithms are trained and evaluated using accuracy, precision, recall, F1-score, and confusion matrices. The system displays prediction results and compares the performance of the implemented models.

C. SOFTWARE REQUIREMENTS

- Programming language: Python
- Libraries: Pandas, NumPy, Scikit-learn
- Visualization: Matplotlib, Seaborn
- Platform: Jupyter Notebook or VS Code
- Hardware: Computer or laptop with at least 4 GB RAM

IV. METHODOLOGY

A. Dataset and Preprocessing

The project report identifies a dataset with 3,772 records and 30 columns. The data include a mixture of numeric and non-numeric values. Preprocessing is therefore an important step before model training. Categorical values must be represented in a form supported by the selected estimators, and missing or inconsistent entries should be handled consistently between training and test data. The report describes preprocessing through the project interface; it does not provide a complete feature-by-feature data dictionary.

B. Classification Algorithms

Decision Tree (DT) learns a sequence of feature-based splits and offers a comparatively interpretable classification path. Random Forest (RF) aggregates the outputs of multiple decision trees to improve stability. Support Vector Machine (SVM) constructs decision boundaries that separate classes in feature space. Multilayer Perceptron (MLP) is a feed-forward neural network capable of learning nonlinear patterns. The implementation runs these algorithms independently to compare their predictive performance.

C. Evaluation Measures

The project evaluates models using accuracy, precision, recall, F1-score, and confusion matrices. Accuracy measures the proportion of correctly classified records. Precision measures the proportion of predicted positives that are correct for a class, while recall measures the proportion of actual members of a class that are correctly identified. The F1-score is the harmonic mean of precision and recall. In a three-class task, precision, recall, and F1-score depend on the class-averaging method; the report does not consistently specify that averaging convention.

For reproducible evaluation, preprocessing should be fitted on training data only, and the test set should remain unseen during model selection. Class-wise metrics and a clearly labelled confusion matrix are particularly important where the number of examples differs among normal, hypothyroid, and hyperthyroid classes.

V. RESULTS

The following results are transcribed from the project report and should be read as the reported experimental outcomes, not as an independent re-run of the models.For the Decision Tree model, the report gives accuracy of about 99.20%, precision of 96.13%, recall of 91.35%, and F1-score of 93.48%. The confusion matrix is described as having most observations on the diagonal, indicating that most records were assigned to their correct class, with a smaller number of off-diagonal errors. The report gives SVM accuracy as 95% and MLP accuracy as 98%. Random Forest is reported at approximately 99% accuracy, and the comparison discussion identifies Random Forest and Decision Tree as the leading models. These results suggest that tree-based methods performed strongly on the dataset used in the project. However, accuracy alone does not establish clinical suitability. The reported class-specific performance, test-set design, possible class imbalance, and validation on data from other sources should be considered before practical deployment. The reported numbers should not be interpreted as a guarantee of performance on new patients.

Algorithm	Reported accuracy
Decision Tree	Approximately 99.20%
Random Forest	Approximately 99%
SVM	Approximately 95%
MLP	Approximately 98%

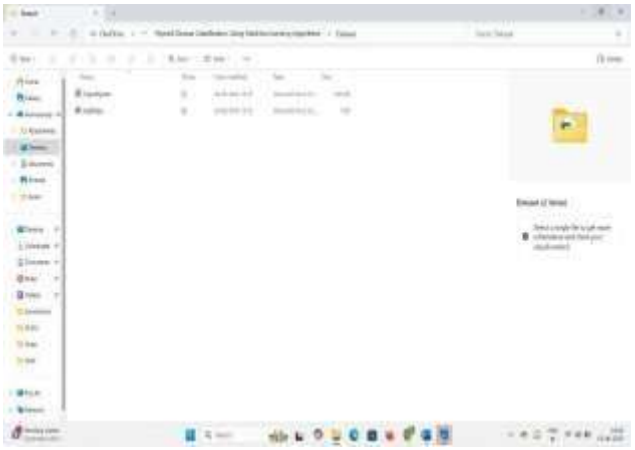


Figure 2:Dataset Directory Structure With Files Of 'Testdata And Hypothyroid Files'

Figure 2 shows the dataset folder containing thyroid-related data files used in the project, including training and testing datasets stored in Excel format for easy access and management. This organized structure helps in the efficient handling and preprocessing of data. The dataset folder is systematically arranged to ensure a smooth workflow during model development and evaluation. It typically includes separate files for training and testing datasets, allowing proper validation of the model's performance.

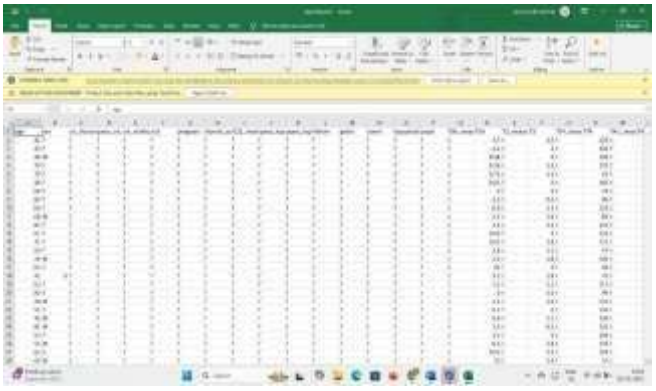


Figure 3: Screenshot of the "hypothyroid " file showing different fields

Figure 3 displays the dataset opened in Excel, where multiple attributes related to thyroid diagnosis are clearly visible. The dataset includes features such as age, sex, medical conditions, and hormone levels like TSH, T3, and T4. Each row represents an individual patient record, while each column corresponds to a specific feature.



Figure 4 : GUI/Main Interface

Figure 4 shows the graphical user interface of the Thyroid Disease Classification system developed using machine learning algorithms. The interface provides multiple buttons for performing different operations such as uploading the dataset, preprocessing the

data, and running various algorithms. It includes options to execute models like Decision Tree, SVM, Random Forest, Naive Bayes, Logistic Regression, KNN, and MLP algorithms.

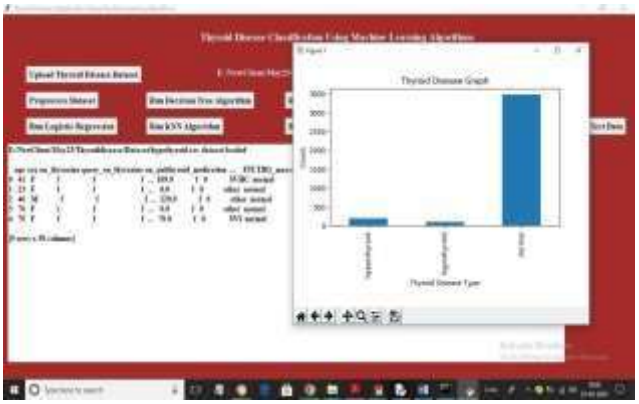


Figure 5: Dataset of Thyroid Disease Classification Using Machine Learning Techniques

Figure 5 shows the output visualization section of the system, where results are displayed after executing the machine learning process. The image appears to contain a mixed or overlapped visual output, indicating multiple layers of data or images combined together. This may represent processed results, intermediate outputs, or a visualization artifact generated by the system. The presence of different textures and regions suggests that the system is handling complex data transformations. It could also indicate an issue in rendering or merging outputs during execution. The figure highlights how the system attempts to present visual results in the output panel.

Overall, it represents the system's capability to generate and display result-oriented visuals. It also indicates that the system may require further optimization in handling and rendering visual outputs. Improving the clarity of such images can enhance user understanding and analysis of results. This figure suggests the importance of proper data visualization techniques in machine learning applications. Clear and well-structured outputs help in better interpretation and decision-making. Overall, it highlights the need for refining the graphical output section to ensure accurate and meaningful result presentation.



Figure 6: Preprocessing of Dataset

Figure 6 shows the thyroid disease dataset was collected and preprocessed before applying machine learning algorithms. The dataset contains 3,772 records and 30 features, including age, sex, thyroid-related symptoms, and laboratory measurements. Data preprocessing involved handling missing values, converting categorical attributes into numerical values, and preparing the data for model training. The preprocessed dataset was divided into training and testing sets, with 80% (3,017 records) used for training and 20% (755 records) used for testing. The prepared dataset was then used to train and evaluate different machine learning algorithms.

.

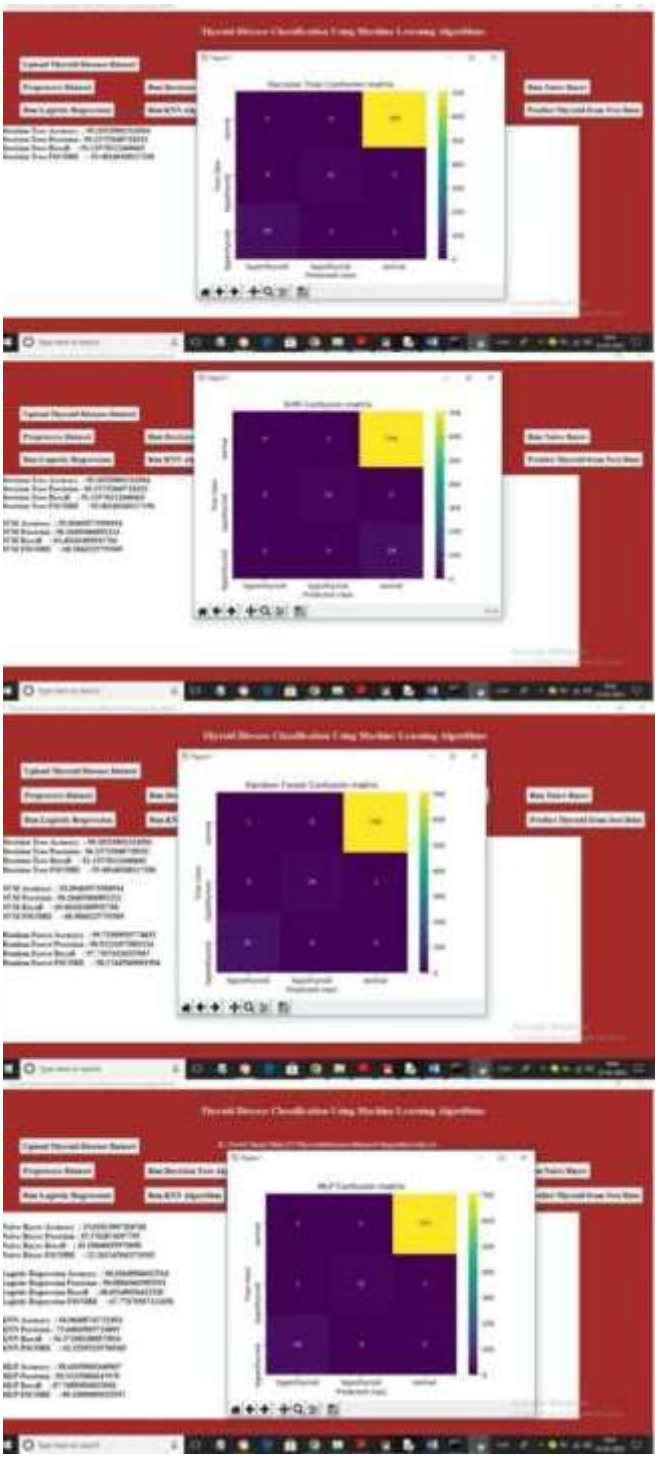


Figure 7: Evaluation Metrics and Confusion Matrix of Decision Tree, SVM, Random Forest, MLP algorithms

Figure 7 shows the confusion matrix of the Decision Tree algorithm used in the thyroid disease classification system. A confusion matrix is used to evaluate the performance of a classification model by comparing the actual classes with the predicted classes. In this figure, the classes considered are hyperthyroid, hypothyroid, and normal. The diagonal values represent correct predictions, while the off-diagonal values indicate misclassifications. From the matrix, most of the values are concentrated along the diagonal, which indicates that the model is performing well.

In the image with SVM we got 95% accuracy and we can see other metrics and confusion matrix graph. In the image, with Random Forest we got 99% accuracy and now click on 'MLP Algorithm'. In above screen MLP got 98% accuracy.

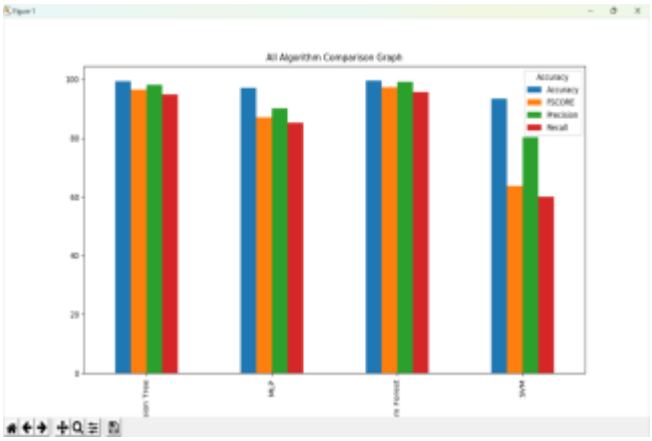


Figure 8: Comparison graph of all Algorithms

Figure 8 illustrates the performance comparison of different machine learning algorithms—Decision Tree, MLP, Random Forest, and SVM—based on evaluation metrics such as Accuracy, F1-Score, Precision, and Recall. From the graph, it is evident that Random Forest achieves the highest performance across most metrics, showing excellent accuracy and balanced precision and recall, making it the most effective model for this dataset.



Figure 9: Prediction of Test Data

Figure 9 shows the prediction output of the thyroid disease classification system using test data. In this interface, after training the machine learning models, new unseen test data is given as input to the system. Each row labeled as "Test Data" represents a patient's medical attributes such as age, hormone levels, and other clinical parameters.

CONCLUSION

In conclusion, the project "Thyroid Disease Prediction Using Data-Driven Machine Learning Algorithms" successfully demonstrates the application of machine learning techniques for predicting thyroid diseases based on patient data. The By applying data-driven machine learning algorithms, the system supports efficient data analysis and provides valuable insights that may assist healthcare professionals in the early identification of thyroid-related conditions. Future enhancements can include incorporating larger and more diverse medical datasets, exploring advanced machine learning algorithms, and improving prediction accuracy. Overall, this project highlights the potential of data-driven machine learning in supporting thyroid disease prediction and contributing to improved healthcare decision-making. However, the system should be used as a supportive tool and not as a replacement for professional medical diagnosis.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the Department of Electronics and Communication Engineering, CMR Technical Campus (Autonomous Engineering College), Hyderabad, Telangana, India, for providing the necessary facilities, infrastructure, and technical support to carry out the project titled "Thyroid Disease Prediction Using Data-Driven Machine Learning Algorithms." The authors would also like to extend their heartfelt appreciation to the faculty members, project supervisor, and technical staff for their continuous guidance, encouragement, and valuable suggestions throughout the development and implementation of this project. Their support and constructive feedback have played an important role in the successful completion of this work.Finally, the authors express their gratitude to everyone who directly or indirectly contributed to the successful completion of this project.

REFERENCES

[1] Azar, at, Hasaean, A.E. and Kim, T. Expert system based on neural fuzzy rules for thyroid diseases diagnosis, Computer Science, Artificial Intelligence Pp. 1-12,2012.

[2] Kales, A. ESTDD: Expert system for thyroid diseases diagnosis, Expert Syst Appl., Vol. 34, No.1, Pp.242–246,2008.

[3] Schuck, "World Health Organization," 2000. [Online]. Available: https://www.who.int/. 2nd International Conference on Physics and Applied Sciences (ICPAS 2021) Journal of Physics: Conference Series 1963 (2021) 012140 IOP Publishing doi:10.1088/1742-6596/1963/1/012140 11

[4] Koura, K., Exarches, T.P., Exarches, K.P., Karamus, M.V. and Fotiadis, D.I. (2015) Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, Vol. 13, Pp.8–17.

[5] Shukla, A. & Kaur, P. (2009). Diagnosis of thyroid disorders using artificial neural networks, IEEE International Advanced Computing Conference (IACC 2009)– Patiala, India, pp 1016-1020.

[6] Aswad, Salma Abdullah, and Eulah Sinus. "Classification of VPN Network Traffic Flow Using Time Related Features on Apache Spark." 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT). IEEE, 2020.

[7] Banu, G. Rasta. "A Role of decision Tree classification data Mining Technique in Diagnosing Thyroid disease." International Journal of Computer Sciences and Engineering 4.11 (2016): 64- 70.H. Mosalli, S. Sanami, Y. Yang, H. G. Yeh & A. G. Aghdam, Dynamic Load Balancing for EV Charging Stations Using Reinforcement Learning and Demand Prediction,arXiv,2025, arXiv:2503.06370

[8] Chindi, Jamil Ahmed, et al. "TDV: Intelligent system for thyroid disease visualization." 2016 International Conference on Computing, Electronic and Electrical Engineering (ICE Cube). IEEE, 2016.