DEFINE RESEARCH JOURNAL
Online Open Access publishing platform for Multidisciplinary Research

Research Article **ISSN 3107 – 7927**

Machine Learning Applications in the Detection and Prognosis of Thyroid Carcinoma for Precision Medicine

Shellymol[1] , Asha Mary Chacko[2]

1-Research Scholar,ShriVenkatrshwara University, Gajraula, UP

Email: Shellymol1986@gmail.com

2- Research Scholar, IES University Bhopal

Email:ashamarychacko1994@gmail.com

Abstract*—Thyroid carcinoma (THCA), a widespread endocrine malignancy, is distinguished by its molecular heterogeneousness and typically a symptomatic early phases, resulting indiagnostic and predictive challenges. Diagnostic methods existing traditionally is not able to detect the complex biological nature of THCA, and thus the necessity for advanced analytical methods is arising. This review emphasizes the key role of bio informatics and machine learning (ML) in the enhancement of THCA detection and prognosis, and for the progress of precision medicine. By integrating high-through put gene expression profiles with ML methods (Logistic Regression, Random Forests, Decision Trees, Naive Bayes, Support Vector Machines, and Artificial Neural Networks), scientists have taken great jumps towards enhanced classification accuracy and predictive ability. These models, validated by measures of accuracy, precision, recall, F1- score, and confusion matrices, confirm to the potential of ensemble methods and neural networks in outperforming traditional methods. Besides, the detection and identification of hub genes via computational and clinical research provide indications towards putative THCA biomarkers, enabling targeted therapeutic approaches. Although these advances mark great jumps, issues remain, including model interpretability, data dimensionality, and the requirement for standard methods. This review will encourage future research to improve computational models and include multi-omics data to overcome current limitations.*

Keywords*: Thyroid carcinoma(THCA), Bioinformatics, Gene expression profiling, Differentially expressed genes(DEGs), Hubgenes, Prognostic biomarkers, Machine learning, Logis- tic Regression, Decision Tree, Random Forest, Support Vector Machine(SVM), Naive Bayes, Artificial Neural Network (ANN), Classificationmetrics(accuracy,precision,recall,F1-score),Confusion matrix, Dimensionality reduction, Principal Component Analysis(PCA), Imputation, Standardization, Ensemble models, Integrated bioinformatics analysis, Microarray/ RNA-Seq, GEO, TCGA, Protein–protein interaction (PPI) network, Gene ontology(GO) analysis, KEGG pathway analysis, Survival analysis(Kaplan–Meier, Cox regression), Immune infiltration, LASSO regression,Weighted geneco-expression network analysis (WGCNA).*

1. INTRODUCTION

Thyroid carcinoma (THCA) is one of the most widespread endocrine cancers.This is a great challenge to clinicians because of its molecular heterogeneousness and frequently early presentation. Conventional diagnostic strategies such as fine needle aspiration (FNA) leads to high rates of false negatives . These constraints lead to needless surgery. This exposes patients to treatment-associated risks and they expose to lifelong hormone therapies. To find a solution to these clinical difficulties, the arrival of precision medicine provides new opportunities for early detection, classification, and prognosis of THCA. Specifically, machine learning (ML)

techniques such as six classifiers evaluated in this study (see Section III-D) help with disease stratification and recurrence prediction. These methods can enhance classification performance by applying to high throughput gene expression data, as quantified by accuracy, precision, recall, F1-score, and confusion matrix statistics. Importantly, ensemble-based classifiers like Random Forest has accuracy rate of up to 98% in detecting subtle disease states like compensated hypothyroidism, while providing balanced performance for multiple classes. Assisting these ML developments in pre-processing complex biomedical data for downstream analysis, bioinformatics pipelines such as dimensionality reduction (e.g., PCA), missing-value imputation, and standardization steps have been found effective. Furthermore, integrated analysis based on gene-set enrichment and protein–protein interaction (PPI) network allow the identification of hub genes involved in THCA pathobiology—that are validated using survival analysis, tissue-based clinical assays. GEO and TCGA databases complement these endeavors by providing large-scale, heterogeneous datasets necessary for successful biomarker discovery. Also this study employs an optimized pipeline-including data preprocessing, PCA-based dimensionality reduction, and comparative evaluation of six ML classifiers on a publicly available thyroid dataset. Random Forest model is found as best for early detection and accurate classification. Solving challenges in ML model's interpretability, class imbalance, standardization among studies shows a good evidence of the potential of intersection of bio-informatics and machine learning . This helps in the advancement of precision diagnostics and disease detection in THCA.

2. LITERATURE REVIEW

Yihang Yuan et al. [1] performed a bioinformatics study of colorectal cancer (CRC) based on GEO datasets GSE17538 andGSE29623, contrasting303 tumor and adjacent normal tissues. Identified 30 significant genes in colorectal cancer (CRC) using WGCNA, reduced to three hub genes (HCLS1,CD48, EVI2B). Low expression genes associated with poor prognosis is validated using TCGA. Demonstrates the value of network-based hub gene discovery.

Han Xue et al. [2] conducted an integrative bioinformatics and experimental research to discover prognostic hub genes of cervical cancer, providing a methodology template for thyroid carcinoma research. 89 overlapping DEGs from GEO datasets were screened in cervical cancer and identified four hub genes (CDC45, GINS2, MCM2, PCNA).Reduction in MCM2 suppressed proliferation which highlights their prognostic role. Provides a pipeline applicable to THCA biomarker research.

Hongrui Zhou et al. [3] conducted an integrative bioinformatics analysis to determine hub genes in colorectal cancer (CRC). They used limma, GO/KEGG analysis, and PPI network analysis and identified ERCC6L ,DSN1, AHCY as hub genes. Silencing DSN1 suggests therapeutic targets.

Yi-Xuan Deng et al. [4] used two GEO datasets (GSE2208 and GSE56815) to compare postmenopausal osteoporosis (PMOP). Each containing 25 low-BMD women and 25 controls. Identified FOS, PTPN6, CTSD as hub genes. These genes are then validated in vivo to strengthen biomarker discovery.

Ting Zhao et al. [5] integrated two GEO datasets (GSE33335andGSE79973)with TCGAgastric cancer data and found 84 common DEGs (34 upregulated, 50 down regulated).Found COL1A1 and COL4A1 as poor-prognosis hub genes. Demonstrates

the power of combining transcript tomic data with survival and protein-level validation.

Samira Nomiri et al. [6] investigated colorectal cancer (CRC) downloading GEO dataset GSE89393 containing 11 paired tumor and normal tissues.Applied WGCNA to CRC and found GUCA2B as a prognostic biomarker. It can be used for RNA-level regulatory network exploration in THCA.

YanzhiGe *et al.* [7] performed a multi-dataset bioinformatics analysis of OA (Osteoarthritis) and RA(Rheumatoid arthritis) synovial tissues, integrating 11 GEO datasets and 265 specimens. They found 10 hub genes enriched in protein synthesis pathwaysthat includes RPS14,RPS6, RPS25, RPL11, SNRPE, RPL27, EEF2,RPS29,RPL10A,andRPL19. Immuno-profiling revealed distinct immune infiltration, showing value of cross-species validation.

Junqiang Xue et al. [8] investigated acute myocardial infarction (AMI) by analyzing GEO data, screening 2,102 DEGs and employing WGCNA to find a notable ”brown” module. Identified ribosomal proteins (RPL9 and RPL26) as hub genes underscoring utility of WGCNA with clinical validation.

Xiao-Qing Lu et al. [9],Combined seven gastric cancer (GC) datasets, identified295DEGs(117upregulated,178 down regulated) and found seven top up regulated hub genes (FN1,HMMR, SPP1, MAD2L1,CCNB1, CXCL8, CCNA2) linked with poor prognosis. Offers template for THCA biomarker discovery.

Yang-Yang Zhou *et al.* [10] used four GEO datasets and conducted an integrative transcriptomic analysis of pancreatic cancer .Constructed 19-gene prognostic signature for pancreatic cancer via WGCNA and Cox regression, showing how integrative pipelines can build predictive models.

Li et al. [11] used GEO2R and DAVID to analyze an a plastic thyroid carcinoma (ATC) and to determine differentially expressed genes (DEGs) and clarify its aggressive biology. Focused on an a plastic thyroid carcinoma and identified cell cycle–related hub genes (TOP2A, CCNB1, CDK1, BUB1), that can be used as targets for precision therapy.

Shen et al. [12] performed integrated bioinformatics analysis on TCGA and GEO datasets to discover prognostic hub genes in papillary thyroid carcinoma (PTC). Identified FN1, SERPINA1, and SLC34A2 as prognostic genes in papillary thyroid carcinoma and is validated by qRT-PCR. This is directly relevant to THCA prognosis modelling.

Wang et al. [13] used WGCNA to discover metastasis biomarkers in differentiated thyroid cancer (DTC).Used WGCNA to link FN1, COL1A1, POSTN to metastasis in differentiated thyroid cancer, implicating ECM and EMT pathways.

Priya and Suganthi [14] contrasted the same six machine learning classifiers used in this study for classifying thyroid disease.Random Forest and ANN performed best. This study uses those.

3. PROPOSED METHODOLOGY

For the development and evaluation of machine learning model to diagnose thyroid disease, following methodology is used. It involves system overview,data acquisition,

data preprocessing, and application of different machine learning algorithms. The ultimate aim is to find the best model using clinical data.

A. System Overview

The proposed system uses a structured machine learning work flow to categorize thyroid conditions with precision and reliability .After complete pre processing(see SectionIII-C), multiple classifiers are then trained and compared to ensure their evaluation is strong. Then their performance assessed through accuracy, precision, recall, F1-score, and confusion matrices. This multi-metrice valuation provides awareness about the strengths and weaknesses of each model, particularly in distinguishing compensated hypothyroid, primary hypothyroid, and negative (healthy) conditions. The overall architecture provides accurate benchmarking while indicating areas for development, enabling the choice of the most suited algorithm for clinical decision support.

B. Dataset Description

This research uses the Thyroid Disease Data dataset from Kaggle [15], which comprises clinical and diagnostic criteria to categorize patients into negative (normal), compensated hypothyroid, and primary hypothyroid classes. The dataset comprises both numerical and categorical variables with some missing values, indicated by "?". The final column shows the thyroid illness status (class label). The dataset served as the basis for training and evaluating machine learning classifiers in the proposed system.

C. Data Preprocessing

This is to confirm that the dataset is machine learning–ready. Using NaN, place holder values such as "?" were replaced and imputed using the mean strategy. Object-type features were converted to numeric format. Standardizationis done with Standard Scaler. All the attributes were normalized. Additionally, to reduce dimensionality ,PCA was applied and this facilitates visualization. For 2D class distribution plots two components were retained.It is critical to elucidate that the PCA transformation was used only for visual ization purposes. All machine learning models were trained and evaluated on the original standardized feature set. This preserves all predictive information for the classification task. To allow impartial evaluation, the dataset was divided into training and testing subsets in the ratio 80:20. Six classification models were tested on this. Finally, this processed dataset is then used for all successive training and validation steps.

D. Machine Learning Implementation

To evaluate diagnostic capability, Six classification models were tested: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Artificial Neural Network (ANN) with a Multilayer Perceptron (MLP) architecture. Their detailed configurations and preprocessing pipeline are described in earlier sections; here we summarize performance outcomes.

Decision tree model achieved 96% accuracy, Logistic Regression model achieved 95% accuracy, Random Forest model achieved 98% accuracy, Support Vector Machine achieved 95% accuracy, Naïve Bayes model achieved 95% accuracy, ANN model achieved 97% accuracy. Here Random Forest model is the best design.it shows strong classification of minority classes. ANN model shows strong

generalization.Naïve Bayes is weak in F1 score. SVM and Logistic regression model struggled with compensated hypo-thyroid. Decision tree has a better accuracy in balance precision and recall.

Overall, Random Forest emerged as the most reliable model, followed by ANN.Hyper parameters were adjusted for optimal performance These results confirm that ensemble and neural architectures are particularly effective for thyroid disorder prediction.

Performance of each model depends on its hyper parameters. Table I outlines the accuracy, precision, recall and F1 score of each model used in this study.

Model	Hyper parameters	Accuracy (%)	Precision (avg)	Recall (avg)	F1-score (avg)
Logistic Regression	max_iter = 500	95	0.95	0.95	0.94
Decision Tree	default	96	0.96	0.96	0.96
Random Forest	nestimators=100,maxdepth=10,minsamplessplit = 4	98	0.98	0.98	0.98
SVM	kernel=rbf, C=1.0, gamma=scale	95	0.94	0.95	0.94
Naïve Bayes	default	95	0.93	0.95	0.94
ANN (MLP)	hidden_layer_sizes=(100,), activation='relu', solver='adam', max_iter=1000	97	0.97	0.97	0.97

TABLEI:SummaryofHyperparameters&PerformanceComparisonofMLModelsonThyroidDataset

4. RESULTS AND DISCUSSION

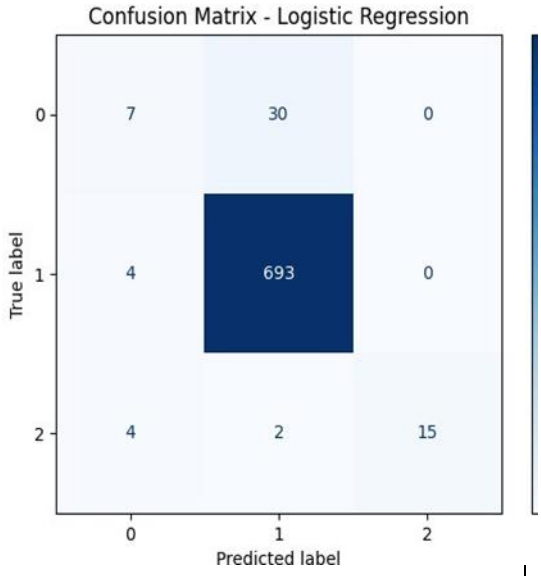
This section presents a comparison of the machine learning classifiers implemented on the thyroid dataset with respect to their accuracy, weighted average precision, recall, and F1- score as presented in Table I. From among all the models, the Random Forest classifier had the best overall performance with an accuracy of 98% and weighted average measures of 0.98 for precision, recall, and F1-score.These findings support its high predictive accuracy and stability across all diagnostic classes, even minority classes like compensated and primary hypo thyroid.Just behind the Artificial Neural Network (ANN) that was trained with a Multilayer Perceptron, with an accuracy of 97% and weighted average precision, recall, andF 1-score of 0.97. This demonstrates the strong ability of ANN to capture intricate, non-linear relationships in the dataset which can therefore be strongly considered for clinical use.

The decision tree model also achieved a balanced average metrics of 96% accuracy, but slightly less than random forest which is 98% accurate. And this model is so simple. This also can be a better choice for diagnostics.

Models like SVM and Logistic Regression both had 95% accuracy with weighted average precision, recall, andF1-scores of 0.94. While their performance was generally strong, the slightly reduced weighted recall and F1-scores indicate that these models might be worse at detecting underrepresented classes. The Naive Bayes model, sharing the same accuracy and weighted averages, is consistent with similar weaknesses likely because of its simplifying assumptions regarding feature independence.

Review of the confusion matrices of all models verified that the majority of classifiers were successful in predicting the majority class (negative),but were more or less successful for the minority classes. Specifically, the Random Forest and ANN models had high recall and precision for all classes, whereas more basic models like Naive Bayes and Logistic Regression had a tendency tom is predict instances of minority classes, which is typical of unbalanced medical datasets.

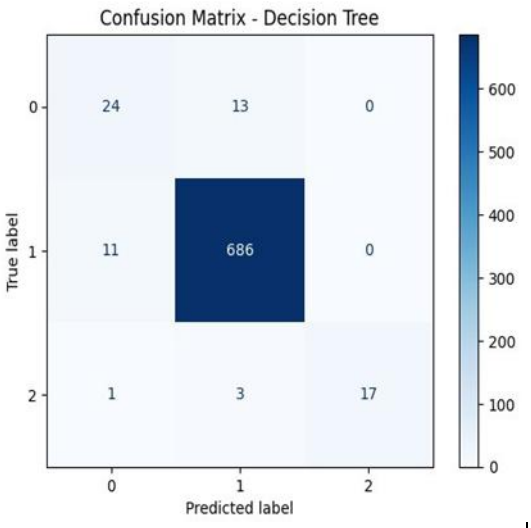
	precision	recall	f1-score	s
compensated_hypothyroid	0.47	0.19	0.27	
negative	0.96	0.99	0.97	
primary_hypothyroid	1.00	0.71	0.83	
accuracy			0.95	
macro avg	0.81	0.63	0.69	
weighted avg	0.93	0.95	0.94	



a. Confusion matrix-Logistic Regression

--- Decision Tree ---

	precision	recall	f1-score	support
compensated_hypothyroid	0.67	0.65	0.66	37
negative	0.98	0.98	0.98	697
primary_hypothyroid	1.00	0.81	0.89	21
accuracy			0.96	755
macro avg	0.88	0.81	0.84	755
weighted avg	0.96	0.96	0.96	755

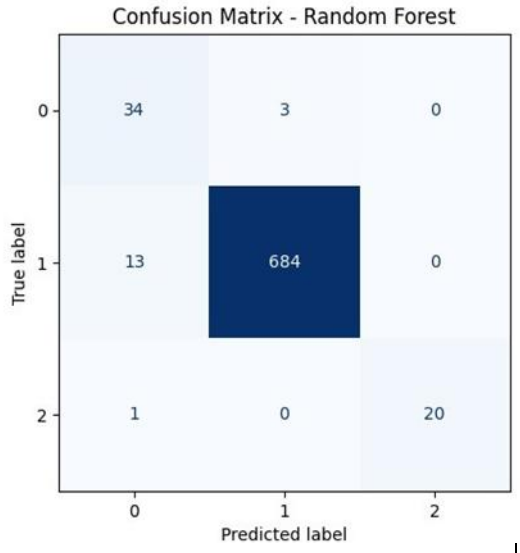


b. Confusion matrix- Decision Tree

Shellymol[1] , Asha Mary Chacko[2]

--- Random Forest ---

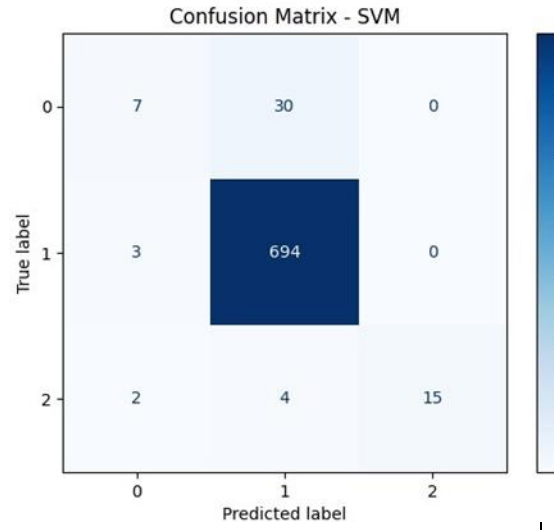
	precision	recall	f1-score
compensated_hypothyroid	0.71	0.92	0.80
negative	1.00	0.98	0.99
primary_hypothyroid	1.00	0.95	0.98
accuracy			0.98
macro avg	0.90	0.95	0.92
weighted avg	0.98	0.98	0.98



c. Confusion matrix- Random Forest

--- SVM ---

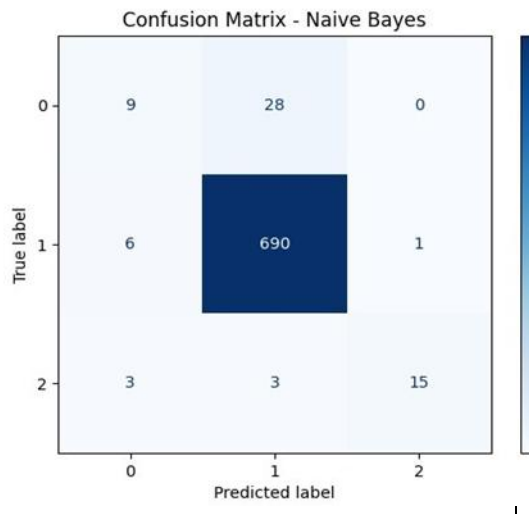
	precision	recall	f1-score	sup
compensated_hypothyroid	0.58	0.19	0.29	
negative	0.95	1.00	0.97	
primary_hypothyroid	1.00	0.71	0.83	
accuracy			0.95	
macro avg	0.85	0.63	0.70	
weighted avg	0.94	0.95	0.94	



d. Confusion matrix- SVM

--- Naive Bayes ---

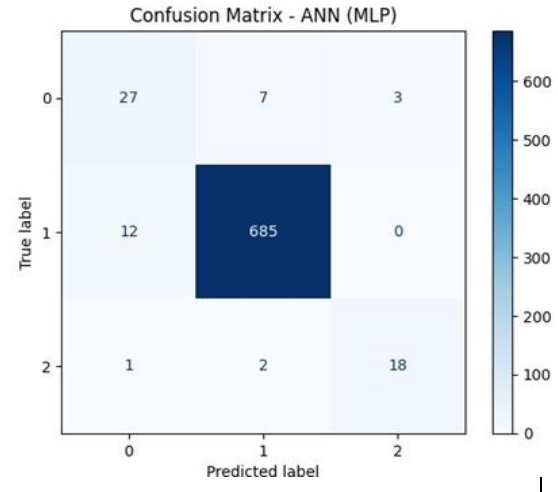
	precision	recall	f1-score	su
compensated_hypothyroid	0.50	0.24	0.33	
negative	0.96	0.99	0.97	
primary_hypothyroid	0.94	0.71	0.81	
accuracy			0.95	
macro avg	0.80	0.65	0.70	
weighted avg	0.93	0.95	0.94	



d. Confusion matrix – Naïve Bayes

--- ANN (MLP) ---

	precision	recall	f1-score	support
compensated_hypothyroid	0.68	0.73	0.70	37
negative	0.99	0.98	0.98	697
primary_hypothyroid	0.86	0.86	0.86	21
accuracy			0.97	755
macro avg	0.84	0.86	0.85	755
weighted avg	0.97	0.97	0.97	755



e. Confusion matrix- ANN

Fig.1:Confusionmatricesofallclassifiersusedforthyroid disease classification

Overall, the Random Forest model was the most accurate and consistent classifier for thyroid classification, closely followed by the ANN model. Both models yielded high accuracy and strong weighted average scores, which make them ideal for being implemented in clinical decision support systems for the diagnosis of thyroid disease.

A. Dimensionality Reduction &Visualization

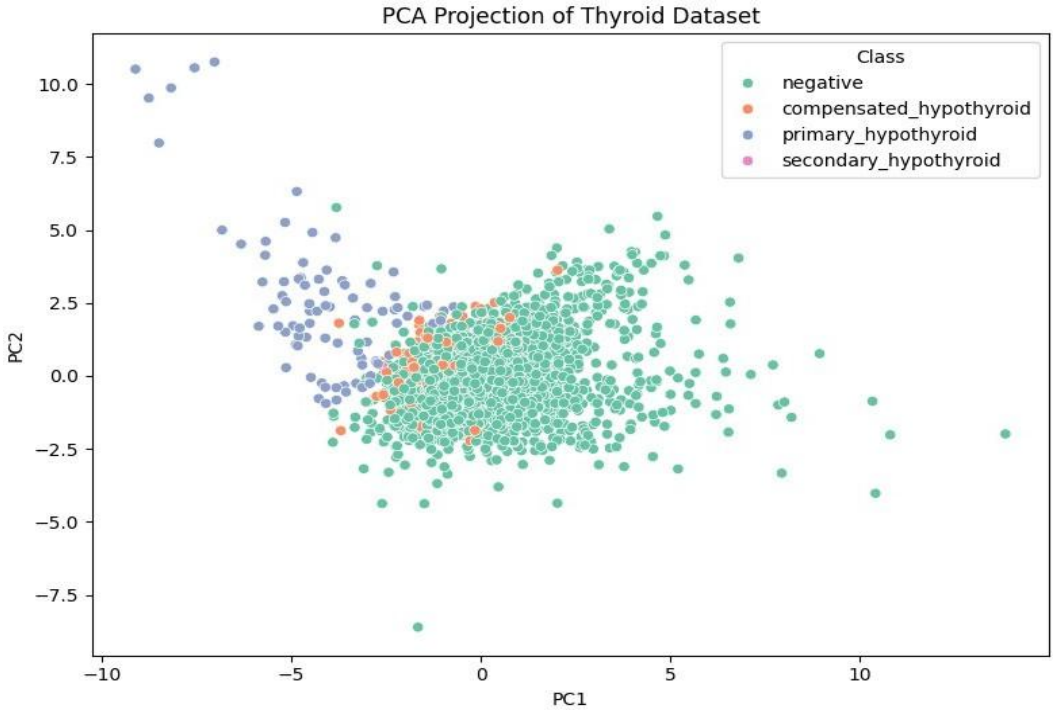


Fig. 2: PCA Projection of Thyroid Dataset across Diagnostic Categories

Figure 2 shows a two-dimensional view of the thyroid dataset projected by Principal Component Analysis (PCA). The points of data in the first two principal components' space (PC1andPC2)are plotted with distinct colors for different di- agnostic classes: negative, compensated hypothyroid, primary hypothyroid, and secondary hypothyroid.

From the plot, one can see that most data points represent the negative class, which is heavily concentrated at the center of the PCA space. The main draw back class consists of a clean cluster primarily in the top left area, while the compensated hypothyroid cases are dispersed and interwoven within the prevailing negative class area. The second hypothyroid classis very sparse, showing class imbalance.

The visual overlap between the negative and compensated classes is consistent with the patterns of confusion found in models like SVM and Naïve Bayes, where the boundary between such classes is less distinct. This analysis also supports the use of Random Forest and ANN models to capture the detailed class differences in the data more accurately.

B. Feature Importance Analysis

Figure 3 depicts the feature importance rankings obtained from the Random Forest model.

This shows TSH (Thyroid Stimulating Hormone) as the most influential feature in predicting thyroid disorders. Other important features are FTI (Free Thyroxine Index), TT4 (Total Thyroxine), and age. T3(Triiodothyronine) and T4U (Thyroxine Utilization)also plays a relevant role in disease prediction. This validates the Random Forest model's ability to highlight biologically meaningful variables. Such visualization helps practitioners to focuson high-impact clinical measurements.

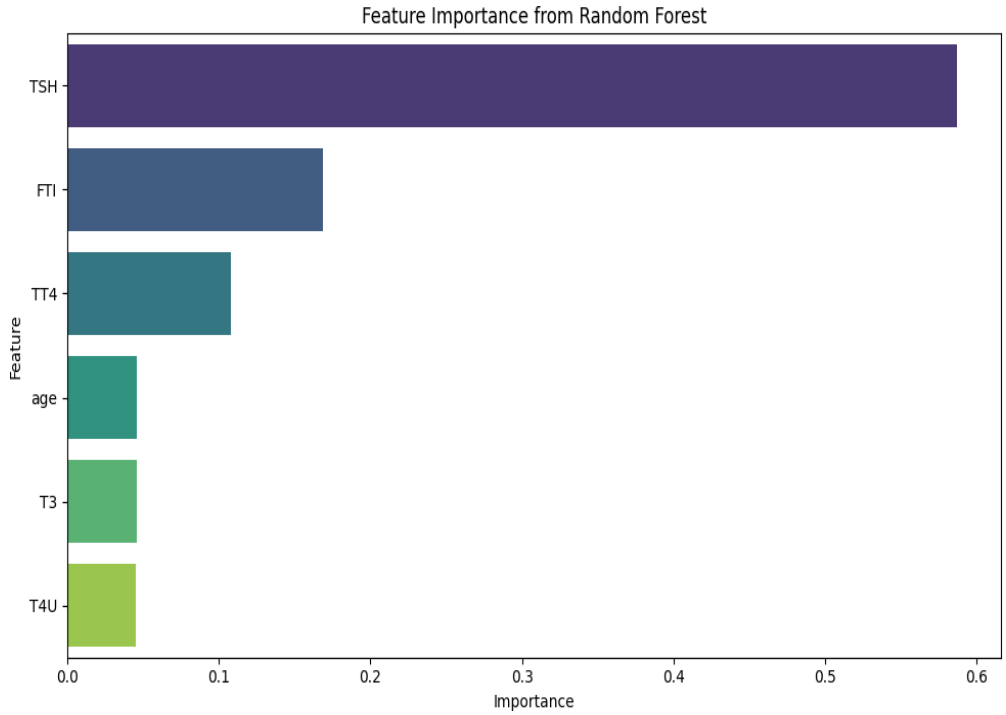


Fig.3:FeatureImportancePlotgeneratedfromtheRandom Forest model

5. DEPLOYMENT FEASIBILITY

Random Forest and Artificial Neural Network (ANN) are the best performing models that can easily be implemented in clinical settings. To make it easier to use in real-world situations, incorporate it into hospital information systems. Using GUI , doctors can enter patient data. Tools such asTkinter, Flask or Dash can be used for this.

Additionally, for the sake of flexibility and availability throughout health care institutions, the whole diagnosis system can be implemented on cloud hosting environments like Amazon Web Services (AWS) or Microsoft Azure. Cloud deployment would facilitate remote diagnostics, large-scale patient volumes calability, and secure data management, and thus the system is ideally suited for both local clinics and vast hospital networks.

6. CONCLUSION

Using real clinical dataset, this study examines the performance of several machine learning models to detect thyroid disorders. This pipe line involves missing value handling, standardization, and PCA-based dimensionality reduction that facilitates improvement in model interpretability and performance.

Six classification models were tested: Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Naive Bayes, and Artificial Neural Network (ANN) with a Multilayer Perceptron (MLP) architecture. Among them, the most powerful was found to be the Random Forest classifier with the best accuracy and performance in precision, recall, and F1-score, especially in identifying all thyroid disorder classes, including the minority ones.

Since ANN(MLP) model shows strong generalization, it can be used as an alternative for rela – world deployment in clinical settings. Conventional models like SVM, Logistic Regression, and Naïve Bayes performed fairly well in absolute accuracy but having class imbalance problems in medical data, so they need to use strong algorithms to deal it. To improve the classification of minority conditions, future efforts should use advanced imbalance and ling methods such as SMOTE, weighting classes, and under sampling. The feature importance analysis shows that TSH, FTI, and TT4 were the strongest predictors in thyroid classification. Class overlap and complexity inthe dataset are visualized through PCA plot.

In general, the results confirm the efficiency of machine learning—specifically ensemble and neural network techniques—as a fruitful method for assisting clinical diagnosis of thyroid diseases. Proper pre processing, careful hype rparameter configuration, and planning of deployment can render such systems as part of hospital processes to help doctors in timely and correct detection of thyroid disease.

7. FUTURE SCOPE

The current work confirms the effectiveness of machine learning algorithms like Random Forest and ANN in thyroid disorder classification with nearly ideal accuracy. Still we can enhance the models in future. A major area for future work is class imbalance. We can use the methods like the Synthetic Minority Oversampling Technique (SMOTE),Adaptive Synthetic Sampling(ADASYN),or class- weighted loss to deal with it. These methods would better increase model sensitivity to minority thyroid disease classes. Also methods like grid search, random search, or Bayesian search can be used for automated hyper parameter tuning to improve performance of models. Further research can use deep learning models—like Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs)—to deal with non-linear connections in clinical data. Using these models in real-world clinical settings and including Explainable AI (XAI) methods will establish confidence, transparency, and acceptance among health care professionals.

REFERENCES

[1] Yuan, Y., Chen, J., Wang, J., Xu, M., Zhang, Y., Sun, P., and Liang, L.,"Identificationofhubgenesincolorectalcancerbyintegratingweighted gene co-expression network analysis and clinical validation in vivo andvitro," Frontiers in Oncology, vol. 10, p. 638, 2020.

[2] Xue,H.,Sun,Z.,Wu,W.,Du,D.,andLiao,S.,"Identification of hub genes as potential prognostic biomarkers in cervical cancer using comprehensive bioinformatics analysis and validation studies," CancerManagement and Research, pp. 117–131, 2021.

[3] Zhou,H.,Yang,Z.,Yue,J.,Chen,Y.,Chen,T.,Mu,T.,Liu,H.,and Bi, X., "Identification of potential hub genes via bioinformaticsanalysis combined with experimental

verification in colorectal cancer,""Molecular Carcinogenesis," vol. 59, no. 4, pp. 425–438, 2020.

[4] Deng,Y.X.,He,W.G.,Cai,H.J.,Jiang,J.H.,Yang,Y.Y.,Dan,Y.R., Luo, H.H., Du, Y., Chen, L., and He, B.C., "Analysis and validation ofhubgenesinbloodmonocytesofpostmenopausalosteoporosispatients,"Frontiers in Endocrinology, vol. 12, p. 815245, 2022.

[5] Zhao,T.,Chen,Z.,Liu,W.,Ju,H.,andLi,F.,"Identificationofhub genes associated with gastric cancer via bioinformatics analysisand validation studies," International Journal of General Medicine, pp.4835–4848, 2023.

[6] Nomiri,S.,Hoshyar,R.,Chamani,E.,Rezaei,Z.,Salmani,F.,Larki,P., Tavakoli, T., Tabrizi, N.J., Derakhshani, A., Santarpia, M., andFranchina,T.,"PredictionandvalidationofGUCA2Basthehub-geneincolorectalcancerbasedonco-expressionnetworkanalysis:in-silicoandin-vivo study," Biomedicine & Pharmacotherapy, vol.147, p. 112691,2022.

[7] Ge,Y.,Chen,Z.,Fu,Y.,Xiao,X.,Xu,H.,Shan,L.,Tong,P.,andZhou,L., "Identification and validation of hub genes of synovial tissue forpatientswithosteoarthritisandrheumatoidarthritis,"Hereditas,vol.158,2021.

[8] Xue,J.,Chen,L.,Cheng,H.,Song,X.,Shi,Y.,Li,L.,Xu,R., Qin, Q., Ma, J. and Ge, J., "The identification and validation of hubgenes associated with acute myocardial infarction using weighted geneco-expression network analysis," Journal of Cardiovascular Developmentand Disease, vol.9, no.1, p.30, 2022.

[9] Lu,X.Q.,Zhang,J.Q.,Zhang,S.X.,Qiao,J.,Qiu,M.T.,Liu,X.R., Chen, X.X., Gao, C., and Zhang, H.H., "Identification of novel hub genesassociated with gastric cancer using integrated bioinformatics analysis,"BMC Cancer, vol.21, no. 1, 2021.

[10] Zhou,Y.Y.,Chen,L.P.,Zhang,Y.,Hu,S.K.,Dong,Z.J.,Wu,M.,Chen, Q.X., Zhuang, Z.Z., and Du, X.J., "Integrated transcriptomic analysisreveals hub genes involved in diagnosis and prognosis of pancreaticcancer," Molecular Medicine, vol.25, 2019.

[11] Li,Y.,Wu,Y.,Yao,H.,Zhu,Y.,Wang,D.andZhang,W.,"Identification of potential biomarkers and biological pathways in an a plastic thyroid carcinoma by bioinformatics analysis," Technology in Cancer Research& Treatment, 2020.

[12] Shen, L., Li, Y., Song, Q., et al., "Identification and validation of hubgeneswithprognosticvalueinpapillarythyroidcarcinomabyintegratedbioinformatics analysis," "Computational and Mathematical Methods inM edicine", 2022.

[13] Wang,J.,Wu,J.,He,Y.,etal.,"Identification of distant meta stasis-related biomarkers in differentiated thyroid cancer using weighted geneco-expression network analysis,"AnnalsofTranslationalMedicine, vol.9, no.13, 2021.

[14] Priya, R. and Suganthi, S., "Thyroid Disease Classification Using Machine Learning Algorithms," in *Proceedings of the International Conference on Physics and Applied Sciences (ICPAS 2021)*, 2021.

[15] Emmanuel Fwerr,"Thyroid DiseaseData,"KaggleDatasets,2023.[Online].Available:https://www.kaggle.com/datasets/emmanuelfwerr/thyroid-disease-data