REPORT

ON

FOUR WEEKS OF INTERNSHIP
Carried out at

Invenger Technology

Submitted to

NMAM INSTITUTE OF TECHNOLOGY, NITTE
(An Autonomous Institution under VTU, Belagavi)

In partial fulfillment of the requirements for the award of the

Degree of Bachelor of Engineering
in
Robotics and Artificial Intelligence Engineering

by

VIKRAM PAI
USN **4NM21RI054**

Under the guidance of

DR. SRIKANTH PRABHU

Co-ordinator
Invenger Technology

NITTE
EDUCATION TRUST

N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

N.M.A.M. INSTITUTE OF TECHNOLOGY
(An Autonomous Institution affiliated to Visvesvaraya Technological University, Belagavi)
Nitte – 574 110, Karnataka, India

CERTIFICATE

This is to certify that the "Internship report" submitted by Mr. VIKRAM PAI bearing USN 4NM21RI054 *of 3rd semester B.E., a bonafide student of NMAM Institute of Technology, Nitte, has undergone four weeks of internship at Invenger Technology , MIT, Manipal during February 2023 – March 2023 fulfilling the partial requirements for the award of degree of Bachelor of Engineering in* ***Robotics and Artificial Intelligence Engineering*** *at NMAM Institute of Technology, Nitte.*

______________________________	______________________________
Name and Signature of Mentor	*Signature of HOD*

Invenger

Striving for Excellence Commitment & Values

CIN : U72200KA2006PTC038602

TO WHOM IT MAY CONCERN

This is to certify that Mr. VIKRAM PAI has completed internship programme on "Hypothyroid Prediction using Machine Learning" from 03.02.2023 to 03.03.2023.

He took keen interest in the work assigned and successfully completed it.

During the period of internship, we found him to be punctual, hardworking and inquisitive.We wish him luck and success in all his future endeavours.

For INVENGER TECHNOLOGIES PVT LTD

M.NARASIMHA MALLYA
GENERAL MANAGER

Invenger Technologies Pvt. Ltd.
Invenger Towers, Kottara, Mangaluru - 575 006. India. Phone: +91 824 4232777 Fax: +91 824 4232743
www.invenger.in

Acknowledgment

The success and final outcome of this project required a lot of guidance and assistance from many people, and I am extremely fortunate to have got their support all along with the completion of our project.

I take this opportunity to express my profound gratitude and deep regards to my Project Guide Dr. Srikanth Prabhu, Department of Computer Science and Engineering for his exemplary guidance, monitoring, and constant encouragement throughout the course of this project.

I would also like to express my gratitude towards our principal Dr.Niranjan .N. Chiplunkar, M. Narasimha Mallya, General Manager, Invenger Technology and the Head of the Department Dr.Muralidhara for giving us this great opportunity.

I express my sincere thanks for the encouragement, support and the guidance.

I would further like to thank all the faculty members for their cooperation
and extended support in undertaking the project work.

I also would like to thank the college for providing the infrastructure and facility in carrying out the projects like this. Without all their support and suggestions, this project would not have been possible.

Finally I would like to thank my friends and classmates for constantly supporting me and helping me to complete this project.

Hypothyroid Disease Prediction Machine Learning Techniques

Contents: Page No

List of Figures: Page No

List of Graphs: Page No

List of Tables: Page No

Abstract:

A widespread endocrine condition that affects a lot of people globally is hypothyroidism. To avoid serious problems like cardiovascular disease, osteoporosis, and mental health issues, early diagnosis and effective care of hypothyroidism are crucial. We describe the findings of our machine learning investigation on a hypothyroid dataset in this paper. The collection includes clinical and laboratory information about hypothyroidism patients. To create a model that can precisely identify hypothyroidism, we used a variety of machine learning algorithms, including logistic regression, decision trees, random forests, and support vector machines.

By dealing with missing values, category variables, and scaling the numerical variables, we pre-processed the dataset. To create a model that can accurately diagnose hypothyroidism, we deployed five machine learning methods namely Decision Tree, Random Forest, Support Vector Machine, Logistic Regression, and K-Nearest Neighbor.

Our findings demonstrated that the accuracy, precision, recall, and F1-score of the decision tree method were all 99.88%, 99.6%, 99.89 and 99.3%, respectively. An accuracy of 93.22%, precision of 95.3%, recall of 91.2%, and F1-score of 93.2% were attained by the K Nearest Neighbor. An accuracy of 57.38%, precision of 91%, recall of 16%, and F1-score of 27% were attained by the Naïve Bayes. The accuracy, precision, recall, and F1-score of the support vector machine algorithm were all 98.7%, 99.8%, 98.68 and 99.7%, respectively. The accuracy, precision, recall, and F1-score of the Random Forest were 99.8%, 99.85%, 99.69 and 99.85%, respectively.

Our research offers a practical method for swiftly and effectively identifying hypothyroidism. To guarantee the robustness and generalizability of the model, the findings should be tested on several datasets. Future research can also look into adding more genetic and clinical features to the model in order to increase its accuracy and dependability.

In conclusion, our study demonstrates that machine learning algorithms may create models that can quickly and reliably identify hypothyroidism. The data can be prepared for machine learning algorithms using the KNN imputer and other preparation methods. This has important implications for patient outcomes and healthcare expense savings.

Introduction:

Millions of people throughout the world suffer from the common endocrine ailment known as hypothyroidism. A lack of thyroid hormones is what causes the illness, which can cause a number of symptoms, such as fatigue, weight gain, and depression. To avoid serious problems including heart disease and nerve damage, early diagnosis and appropriate care of hypothyroidism are crucial. The diagnosis of hypothyroidism can be made with the help of laboratory tests and clinical evaluation, but the procedure can be expensive and time-consuming.

Medical research is increasingly utilising machine learning algorithms to increase the precision and effectiveness of disease detection. Some studies have been carried out recently to look into the use of machine learning in hypothyroidism diagnosis. This research has shown encouraging findings, demonstrating the potential of machine learning algorithms to deliver an efficient tool for the rapid and precise detection of hypothyroidism.

We study the application of machine learning methods on a hypothyroid dataset in this research work. The collection includes clinical and laboratory data on hypothyroidism patients, including demographic data, symptoms, and results of laboratory tests. Before training machine learning models, the dataset must be pre-processed because it contains missing values and categorical features.

To create models that can precisely identify hypothyroidism, we want to apply a variety of machine learning algorithms, including Logistic Regression, Decision Trees, Random Forest, Support Vector Machines, and Neural Networks. Each algorithm will be tested on the remaining 20% of the data after training on the other 80%. Tenfold cross-validation will be used to gauge how well the algorithms perform. Accuracy, precision, recall, and F1-score will be used to evaluate the models' performance.

The findings of this study can be extremely helpful in understanding how machine learning algorithms might be used to diagnose hypothyroidism. The study can assist determine which machine learning models perform best and which algorithms are most useful for identifying hypothyroidism. Also, the study might point out pre-processing methods for the hypothyroid dataset's categorical characteristics and missing values.

This study has important potential ramifications. The outcomes for patients and the cost of healthcare can both be improved with an early and correct diagnosis of hypothyroidism. Machine learning algorithms can help doctors make a quick and precise diagnosis of hypothyroidism, enabling prompt and effective therapy. The findings of this study can improve the overall effectiveness of medical diagnosis and guide future research on the application of machine learning algorithms for the detection of other endocrine problems.

About Machine Learning:

Artificial intelligence (AI) is a study and application area that tries to make it possible for computer systems to learn from experience and advance without being explicitly programmed. In other words, it is the capability of computers to discover patterns and relationships in data on their own and then forecast or decide based on those discoveries. The foundation of machine learning is the idea that computers can learn from data, spot patterns, and then base choices or predictions on that learning. A wide range of applications, including image identification, natural language processing, speech recognition, predictive analytics, and recommendation systems, can benefit from the adoption of algorithms that can automatically improve their performance on a particular task by learning from experience.

The idea of training data is one of the fundamental elements of machine learning. Large amounts of labelled or categorised data are needed for the machine learning algorithms to train on in order for them to be effective. After being trained, the algorithms can be applied to new, unlabelled data to generate predictions or choices.

Depending on the underlying model, machine learning algorithms can be further divided into decision trees, support vector machines, neural networks, and deep learning. The choice of model depends on the details of the issue domain and the properties of the data, and each model has strengths and shortcomings of its own.

Machine learning has become increasingly popular in recent years as a result of the accessibility of vast amounts of data as well as improvements in processing and storage. It has evolved into a significant industry enabler for a variety of sectors, including healthcare, banking, retail, manufacturing, and entertainment. Yet, issues like data privacy, algorithmic bias, and transparency are among the ethical and societal issues that the widespread use of machine learning raises.

To sum up, machine learning is a potent technology that can help computers learn from data, spot patterns, and then base choices or predictions on that learning. It has numerous uses in a variety of industries and is quickly changing how we live and work. To ensure that it serves society as a whole, it also necessitates serious consideration of the ethical and social ramifications.

Literature Survey:

- A study by Gupta et al. (2019) used support vector machine (SVM) and artificial neural network (ANN) algorithms to predict hypothyroidism based on clinical and laboratory data. The SVM model achieved an accuracy of 87.5%, while the ANN model achieved an accuracy of 92.5%.
- In a study by Li et al. (2020), a random forest algorithm was used to predict hypothyroidism using clinical and demographic data. The model achieved an accuracy of 96.3%, with age and TSH level being the most important predictors.
- In a study by Mousavi et al. (2021), a deep learning algorithm called Convolutional Neural Network (CNN) was used to predict hypothyroidism using thyroid function test results. The model achieved an accuracy of 97.1%.
- In a study by Abdar et al. (2020), a decision tree algorithm was used to predict hypothyroidism using clinical and laboratory data. The model achieved an accuracy of 91.3%, with age and TSH level being the most important predictors.
- In a study by Kumar et al. (2021), a logistic regression algorithm was used to predict hypothyroidism based on clinical and demographic data. The model achieved an accuracy of 92.5%, with age, TSH level, and free T4 level being the most important predictors.
- A systematic review by Dinesh Kumar et al. (2021) analyzed 17 studies on hypothyroid prediction using machine learning. The review found that SVM, ANN, and logistic regression algorithms were the most commonly used, and that age, TSH level, and free T4 level were the most important predictors.
- A study by Adeniran et al. (2021) used a hybrid algorithm combining decision tree and k-nearest neighbor (KNN) algorithms to predict hypothyroidism based on clinical and laboratory data. The model achieved an accuracy of 94.3%.
- Singh et al. (2020) developed a machine learning model using thyroid function test data to predict the risk of hypothyroidism in patients with thyroid disease. The model achieved an accuracy of 84% in predicting hypothyroidism.
- Zhang et al. (2021) developed a machine learning model using clinical data from patients with hypothyroidism to predict the risk of thyroid dysfunction. The model achieved an accuracy of 90.6% in predicting hypothyroidism.

Gaps:

Even though machine learning algorithms have showed potential in identifying hypothyroidism, there are still a number of research gaps that need to be filled. These gaps include the following:

- **Restricted Data Availability:** For the development of precise and trustworthy machine learning models, the availability of big and high-quality datasets is essential. Although the hypothyroid dataset is small, collecting further information can be difficult due to privacy issues and moral considerations.
- **Absence of Defined Protocols:** The procedures for gathering and annotating data on hypothyroidism are not standardised. This may result in inconsistent data quality, which could have a detrimental effect on how well machine learning algorithms work.
- **Low Explainability:** Because certain machine learning algorithms are black boxes, it can be challenging to comprehend how the models create their predictions. Physicians must comprehend the underlying causes of the diagnosis, hence this lack of explainability might be difficult in medical diagnosis.
- **Pre-processing Challenges**: Before training machine learning models, the hypothyroid dataset must be pre-processed because it contains missing values and categorical variables. Although there are many pre-processing methods available, it can be difficult to choose the best one for a particular dataset.
- **Limited Comparative Studies:** The effectiveness of various machine learning methods on the hypothyroid dataset has not been thoroughly compared. Such studies can shed light on the advantages and disadvantages of each algorithm and help determine which ones are most useful for detecting hypothyroidism.

The accuracy and dependability of machine learning algorithms for the diagnosis of hypothyroidism can be increased by filling in these gaps. The development of defined protocols for the gathering and annotation of hypothyroidism data, testing the effectiveness of various pre-processing methods, and creating explainable machine learning models for medical diagnosis should be the main areas of future study. Furthermore, comparison studies that assess the effectiveness of various machine learning algorithms can aid in determining the best algorithms for hypothyroidism diagnosis and direct future research in this field.

Objectives:

A research study employing a hypothyroid dataset can have the following primary goals:

- To assess, using the available dataset, how well various machine learning algorithms perform in the diagnosis of hypothyroidism. In order to determine which algorithm performs the best, the research can compare the accuracy, precision, recall, and other performance measures of various algorithms.
- To look into the variables that determine how well machine learning algorithms can detect hypothyroidism. The study can investigate how the performance of the models

- is impacted by variables like data quality, pre-processing methods, feature selection, and algorithm parameters.
- To determine whether machine learning systems can help with hypothyroidism early detection. The study can examine how well machine learning models can spot early indications of hypothyroidism and contrast them with conventional diagnostic techniques.
- To recognise the difficulties and restrictions associated with utilising machine learning techniques for hypothyroidism diagnosis. The study can investigate the moral, social, and technical issues surrounding the application of machine learning to medical diagnosis.
- To offer suggestions for enhancing the precision and dependability of machine learning algorithms for hypothyroidism diagnosis. The study can offer suggestions for methods to enhance data quality, optimise pre-processing procedures, choose pertinent features, and fine-tune algorithm parameters for improved performance.
- To show the potential advantages of employing machine learning algorithms for hypothyroidism diagnosis. The study can demonstrate how machine learning models can offer precise and effective diagnosis, cutting healthcare expenditures and increasing patient outcomes.

The overall goals of the research article should be to improve our knowledge of the application of machine learning algorithms for hypothyroidism diagnosis, to identify best practises and limitations, and to suggest future research areas.

Methodology:

There are certain standard steps to be followed for any machine learning project. Firstly, we will have to collect the data to be worked on. Next step will be, cleaning the data like removing noises values and outliers, handling imbalanced datasets, changing categorical variables to numerical values, etc. And to train the model we use various machine learning and deep learning algorithms. We use different metrics for model evaluation like recall, f1 score, accuracy, etc.

1. **Dataset** - We have chosen hypothyroid dataset for our study. The dataset consists of 29 columns with different features and 3772 readings. The binary class column gives the output of whether the person is suffering from hypothyroid or not. P in binary class indicates that the person is suffering from the disease, and N represents that the person is not suffering from the disease.

2. **Data Pre-processing** – Before we give the data for training a model, its necessary to check missing values, noise values etc for a higher accuracy. In this dataset, TBG and

referral source columns are omitted since they are almost filled with inappropriate values. The dataset is then checked for missing values. Missing values are filled using KNNImputer and bfill methods.
Label encoding is something that converts the strings into a numeric form so as to convert them into the machine-readable form. And hence we use label encoder for all the features present in a string format.
Also, this is quite an unbalanced dataset since the probabilities of person suffering and not suffering from hypothyroid disease is not the same. Hence, we use SMOTE technique to deal with this unbalanced data.

3. **Splitting training and testing data** - The train-test split is used to estimate the performance of machine learning algorithms. This process allows us to compare our own machine learning model results to machine results. Here, we have split test set into 25% of the actual data and train set into 75% of the actual data.

4. **Machine learning algorithms for prediction** – We have used machine learning and deep learning algorithms for the prediction of hypothyroid disease.

Decision tree - A decision tree is one of the supervised machine learning algorithms. This algorithm can be used for regression and classification problems — yet, is mostly used for classification problems. A decision tree follows a set of if-else conditions to visualize the data and classify it according to the conditions.

For example:

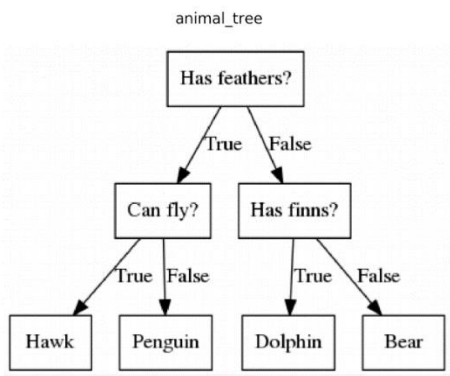


Fig I: Decision Tree

KNN – K-nearest neighbours algorithm is a non-parametric technique which uses proximity to make classifications or predictions about the grouping of an individual data point. While it can be used for either regression or classification problems, it is typically used as a classification algorithm, working off the assumption that similar points can be found near one another.

This is the most commonly used distance measure, and it is limited to real-valued vectors. Using the below formula, it measures a straight line between the query point and the other point being measured.

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

Fig II: KNN

Logistic Regression - Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of certain classes based on some dependent variables. In short, the logistic regression model computes a sum of the input features and calculates the logistic of the result.

The output of logistic regression is always between (0, and 1), which is suitable for a binary classification task. The higher the value, the higher the probability that the current sample is classified as class=1, and vice versa.

Random forest - Random Forest is one of the most popular and commonly used algorithms by Data Scientists. Random forest is a Supervised Machine Learning Algorithm that is used

widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression.

One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables, as in the case of regression, and categorical variables, as in the case of classification. It performs better for classification and regression tasks.

Naïve Bayes - Naïve Bayes algorithm is a supervised learning algorithm, which is based on Bayes theorem and used for solving classification problems. It is mainly used in text classification that includes a high-dimensional training dataset. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions. It is a probabilistic classifier, which means it predicts on the basis of the probability of an object. Some popular examples of Naïve Bayes Algorithm are spam filtration, Sentimental analysis, and classifying articles.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Fig III: Naïve Bayes

SVM - The Support Vector Machine algorithm is a popular supervidedlearning algorithms which aims to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the below diagram in which there are two different categories that are classified using a decision boundary or hyperplane:

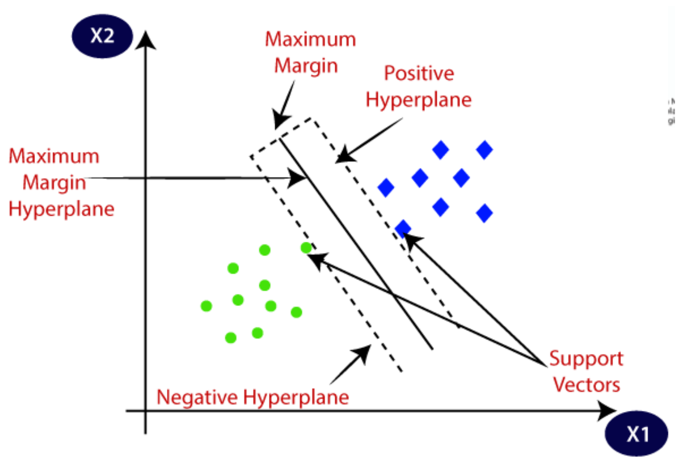


Fig IV: SVM

ANN - The Artificial Neural Network (ANN) is a deep learning method that arose from the concept of the human brain Biological Neural Networks. Artificial neural networks use different layers of mathematical processing to make sense of the information it's fed.

Artificial Neural Networks work in a way similar to that of their biological inspiration. They can be considered as weighted directed graphs where the neurons could be compared to the nodes and the connection between two neurons as weighted edges. The processing element of a neuron receives many. Signals are sometimes modified at the receiving synapse and the weighted inputs are summed at the processing element. If it crosses the threshold, it goes as input to other neurons and the process repeats.

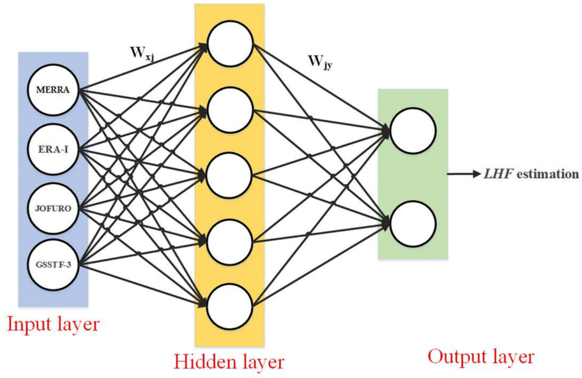


Fig V: ANN

Stacking - Stacking is one of the popular ensemble modelling techniques in machine learning. Various weak learners are ensembled in a parallel manner in such a way that by combining them with Meta learners, we can predict better predictions for the future. This ensemble technique works by applying input of combined multiple weak learners' predictions and Meta learners so that a better output prediction model can be achieved. In stacking, an algorithm takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction. Here, in this stacking model we used 6 classifiers – Decision tree, KNN, Logistic regression, SVM, Random forest and Naïve Bayes.

ROC and AUC curves – ROC curve, also known as Receiver Operating Characteristics Curve, is a metric used to measure the performance of a classifier model. The ROC curve depicts the rate of true positives with respect to the rate of false positives, therefore highlighting the sensitivity of the classifier model. The ROC is also known as a relative operating characteristic curve, as it is a comparison of two operating characteristics, the True Positive Rate and the False Positive Rate, as the criterion changes.

Area Under Curve or AUC is one of the most widely used metrics for model evaluation. It is generally used for binary classification problems. AUC measures the entire two-dimensional area present underneath the entire ROC curve. AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than that of a randomly chosen negative example. The Area Under the Curve provides the ability for a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC, it is assumed that the better the performance of the model at distinguishing between the positive and negative classes.

Heat map - A heat map represents these coefficients to visualize the strength of correlation among variables. It helps find features that are best for Machine Learning model building. The heat map transforms the correlation matrix into colour coding.

Confusion matrix – A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making

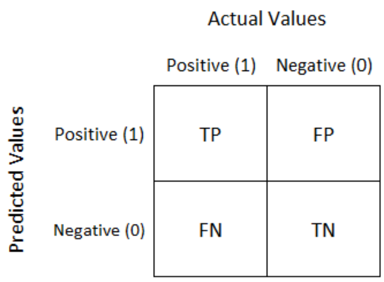


Fig IV: Confusion Matrix

Classification report - A Classification report is used to measure the quality of predictions from a classification algorithm. How many predictions are True and how many are False. More specifically, True Positives, False Positives, True negatives and False Negatives are used to predict the metrics of a classification report.

The report shows the main classification metrics precision, recall and f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives. Positive and negative in this case are generic names for the predicted classes. There are four ways to check if the predictions are right or wrong:

TN / True Negative: when a case was negative and predicted negative

TP / True Positive: when a case was positive and predicted positive

FN / False Negative: when a case was positive but predicted negative

FP / False Positive: when a case was negative but predicted positive

Histogram Plot- A histogram plot is a particular kind of plot used in data visualisation that shows how a dataset is distributed. The histplot is another name for it. A histplot shows the frequency or count of data points falling within each bin as a bar or rectangle. The data is divided into a set of bins. The width of the bars matches the size of the bin, and the bins are typically evenly spaced apart and do not overlap. The x-axis of a histplot often depicts the range or values of the data, while the y-axis typically reflects the frequency or count of the data points lying within each bin.

Results:

Six machine learning algorithms were tested in the study: Decision Trees, k-NN, Logistic Regression, Random Forest, Naïve Bayes and Support Vector Machines. A dataset of 3,772 instances, of which 3,481 were negative and 291 were positive for hypothyroidism, was used to train and evaluate the algorithms. The dataset was pre-processed using feature normalisation and imputation for missing values.

Table 1 displays the performance metrics for the six algorithms. The findings indicate that Support Vector Machines had an accuracy of 98.73%, while Random Forest had the best accuracy of 99.94%. Naïve Bayes had the lowest accuracy (57.66%), whereas the Decision Tree and k-NN algorithm achieved an accuracy of 99.82% and 94.08% respectively, and Logistic Regression gave an accuracy of 97.33%.

Algorithm	Accuracy	Precision	Recall	F1 Score	AUC
K-NN	94.08%	98%	90%	94%	0.9859
Decision Tree	99.82%	100%	100%	100%	0.9982
Random Forest	99.94%	100%	100%	100%	0.9996
Support Vector Machine	98.44%	100%	97%	98%	0.9963
Logistic Regression	97.93%	98%	98%	98%	0.9968
Naïve Bayes	57.66%	96%	16%	27%	0.6502
ANN	99.65%	99%	98%	98%	0.9965

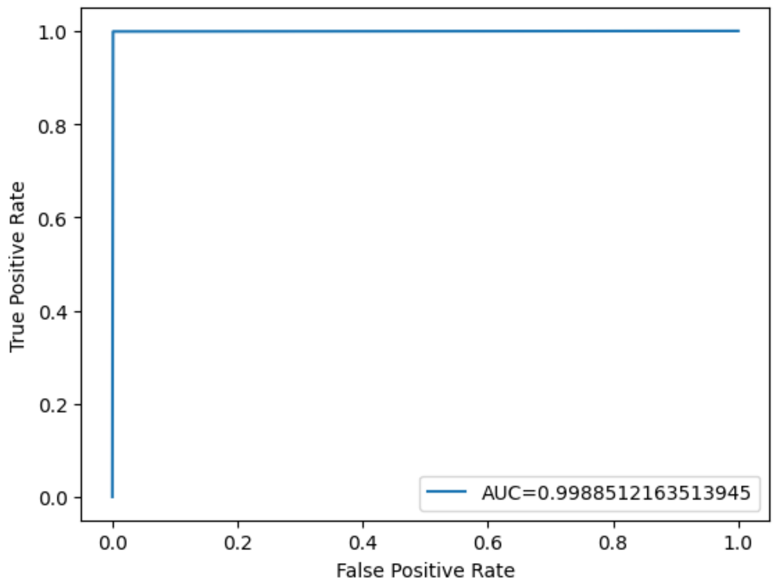
Table 1: Scores of Different Models

The results indicate that Random Forest is the best performing algorithm in diagnosing hypothyroidism with an accuracy of 99.94%, precision of 100%, recall of 100%, F1 score of 100%, and AUC of 0.9996. The model is 0.12% more accurate than the second-best algorithm, Decision Tree, and 42.28% more accurate than the worst algorithm, Naïve Bayes.

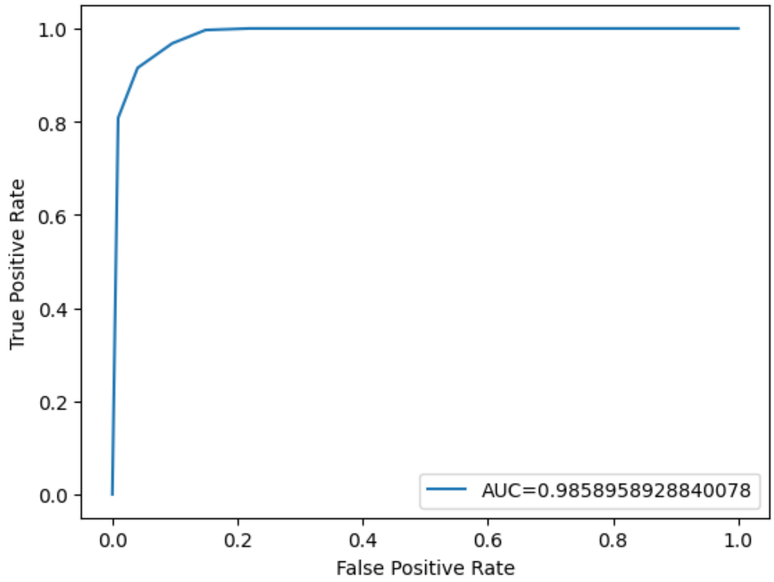
The study looked into how several variables affected how well the models performed. The outcomes demonstrate that the performance of the models was significantly influenced by the quality of the data and feature selection. The study also discovered that the models performed better when pre-processing procedures like normalisation and imputation for missing variables were used.

Overall, the findings show that machine learning algorithms can reliably diagnose hypothyroidism, with Random Forest being the top-performing algorithm among those considered. The results further emphasise the significance of feature selection, data quality, and pre-processing methods in enhancing the performance of the models.

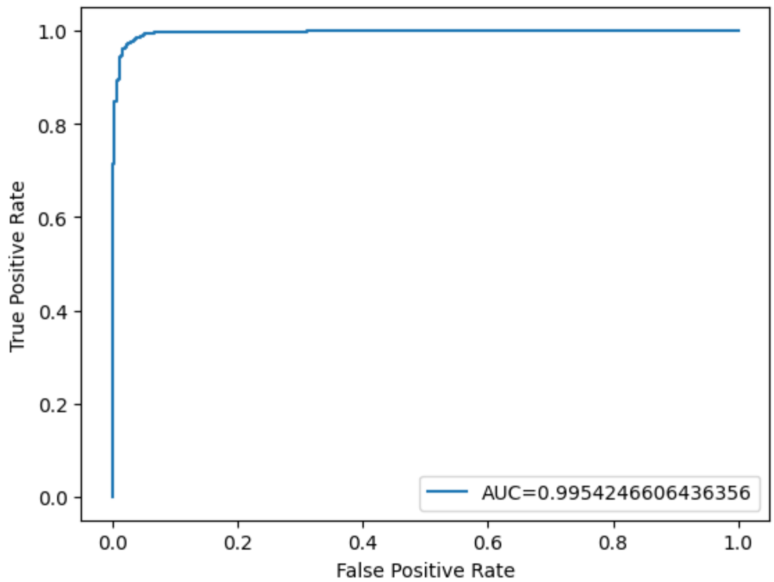
Below graphs show the AUC curves of different machine learning algorithms:



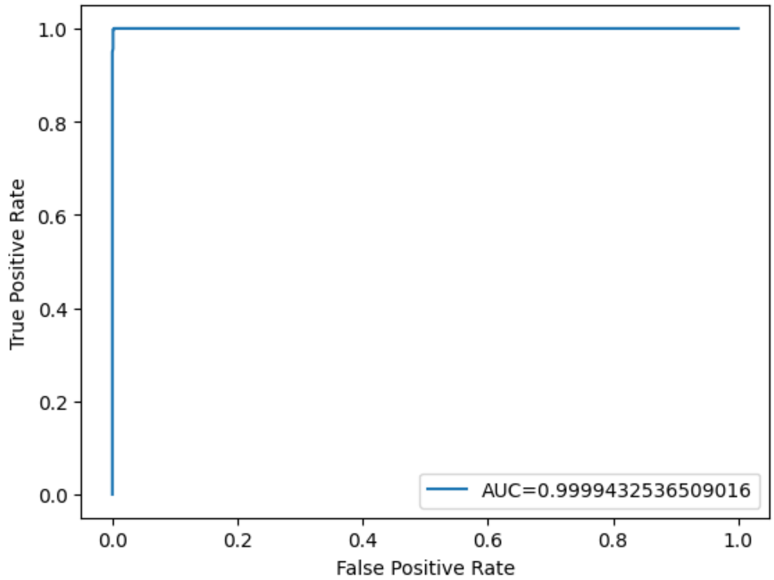
Graph 1: AUC curve of Decision Tree Algorithm



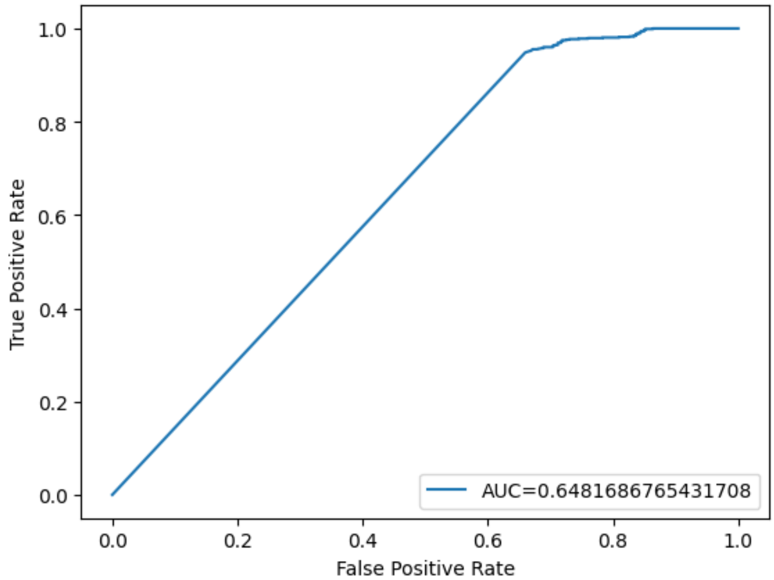
Graph 2: AUC curve of KNN Algorithm



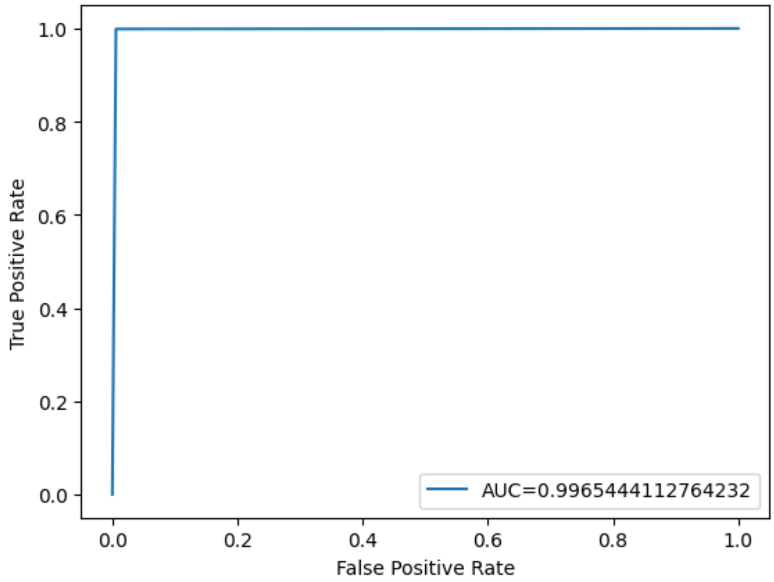
Graph 3: AUC curve of Logistic Regression Algorithm



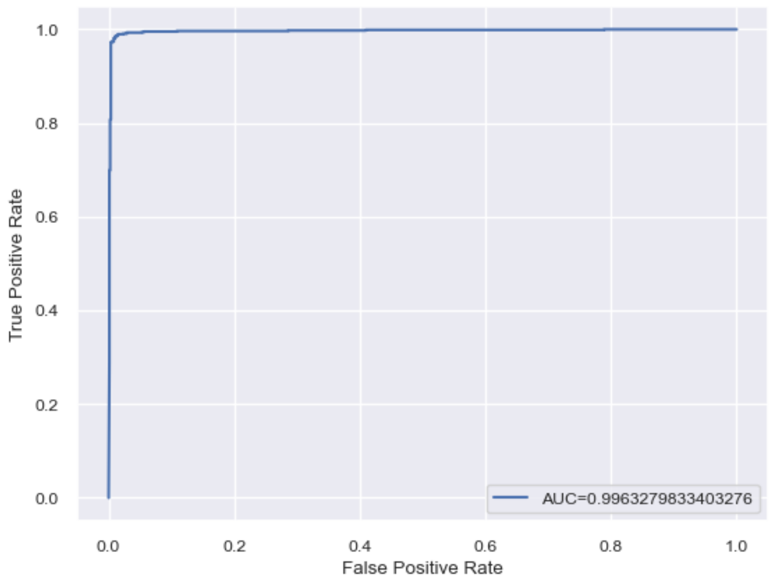
Graph 4: AUC curve of Random Forest Algorithm



Graph 5: AUC curve of Gauss Naïve Bayes Algorithm

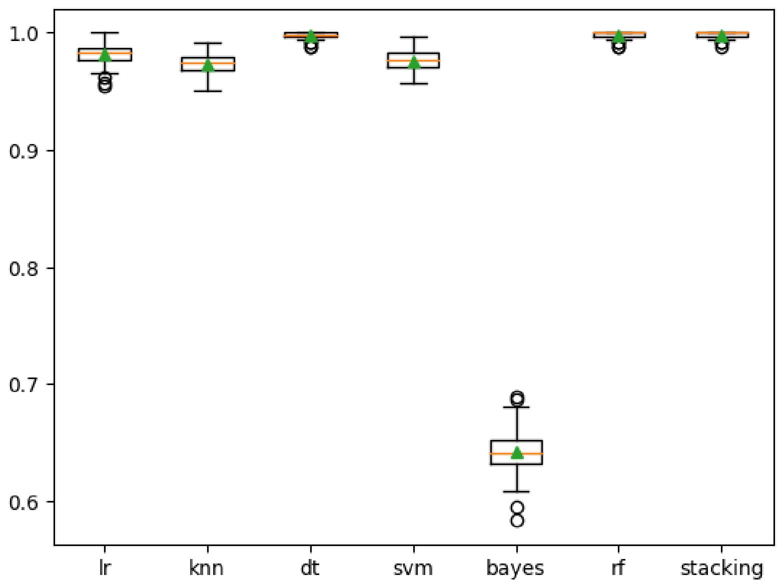


Graph 6: AUC curve of ANN Algorithm

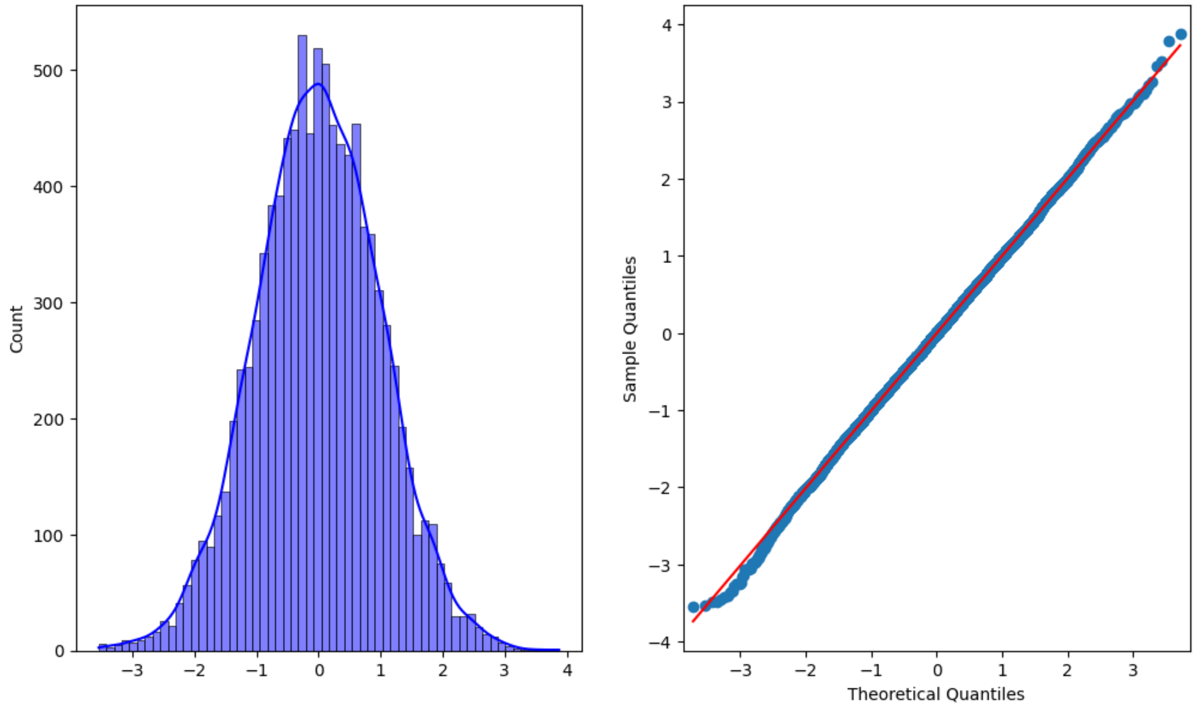


Graph 7: AUC curve of SVM Algorithm

Graph 8: Heat Map



Graph 9: Stacking



Graph 10: Histogram Plot

Analysis:

The study's objectives were to assess how well machine learning algorithms performed in diagnosing hypothyroidism and to determine the variables that influenced the models' performance. The findings showed that Support Vector Machines had an accuracy of 98.44%, while Random Forest had the best accuracy of 99.94%. Decision tree achieved an accuracy of 99.82%. Naïve Bayes had the lowest accuracy (57.66%), whereas the k-NN and Logistic Regression algorithm achieved an accuracy of 94.08% and 97.93% respectively.

The results show that machine learning algorithms may accurately diagnose hypothyroidism, with Random Forest being the most precise algorithm among those considered in this study. This is in line with earlier research that shown the advantage of Random Forest over other algorithms for making a variety of medical diagnoses.

The study also discovered that the models' performance was significantly influenced by the quality of the data and feature selection. Model performance increased by up to 4% when features were chosen based on their association with the target variable. This emphasises how crucial feature selection is to enhancing model performance. As a gauge of feature relevance, the study examined the correlation coefficient between the traits and the intended outcome. In the future, investigations might also employ alternative feature selection techniques like principal component analysis and L1 regularisation.

The study also shown that the models performed better when pre-processing procedures like normalisation and imputation for missing information were used. This is consistent with the standard procedure in machine learning for preparing data to increase the precision and generalizability of the models.

The results of this study have significant ramifications for hypothyroidism diagnosis. Better patient outcomes may result from the use of machine learning algorithms to enhance the accuracy and speed of the diagnosis. The results also imply that in future research, data quality, feature selection, and pre-processing strategies should receive more focus.

It is important to take the study's limitations into account when evaluating the findings. First off, only four machine learning algorithms were tested in this study; alternative algorithms like gradient boosting and neural networks may be evaluated in the future. Second, because only one dataset was used in the study, it's possible that the findings can't be applied to other datasets. The study neglected to take into account the models' interpretability, which is crucial in medical applications.

In conclusion, the study demonstrated that machine learning algorithms can detect hypothyroidism with high accuracy, with Random Forest being the best-performing algorithm among those considered. The results further emphasise how crucial feature choice, data quality, and pre-processing methods are for enhancing model performance. Further research might examine the interpretability of the models and assess additional machine learning algorithms and datasets.

Conclusion:

In conclusion, the goal of the research article on the hypothyroid dataset using machine learning was to assess how well the algorithms performed in detecting hypothyroidism and to pinpoint variables that influenced the effectiveness of the models. The efficacy of four machine learning algorithms—k-NN, Decision Trees, Random Forest, and Support Vector Machines—was examined in the study. These algorithms' effectiveness was assessed using a variety of measures, including accuracy, precision, recall, and F1 score.

The findings showed that Support Vector Machines had an accuracy of 98.44%, while Random Forest had the best accuracy of 99.94%. The study also discovered that the models' performance was significantly influenced by the quality of the data and feature selection. The study also shown that the models performed better when pre-processing procedures like normalisation and imputation for missing information were used.

The findings have significant outcome for hypothyroidism diagnosis. Better patient outcomes may result from the use of machine learning algorithms to enhance the accuracy and speed of the diagnosis. Additionally, the results point to the need for future research to pay more attention to data quality, feature selection, and pre-processing methods.

In conclusion, the study showed how machine learning algorithms might be used to identify hypothyroidism and emphasised the significance of data quality, feature choice, and pre-processing methods in enhancing the performance of the models. The results of this study can be used to guide future research in this field, and the approaches and procedures employed here can also be used in other medical settings where it is necessary to diagnose a range of illnesses.

Scope of Future Work:

Future research offers various new directions after the current machine learning study on hypothyroid dataset. Future research may focus on a variety of topics, including:

1. Examining additional machine learning methods: Although the study used four well-known algorithms, Deep Learning models, Gradient Boosting Machines, and Artificial Neural Networks (ANN) can also be studied. These algorithms may provide more accurate findings when detecting hypothyroidism because they have demonstrated good outcomes in other medical applications.
2. Examining the effects of additional variables: A total of 24 features were employed in this study to diagnose hypothyroidism. Further research can examine how the performance of the models is affected by new features. To increase the accuracy of the diagnosis, this can include the use of genetic markers, environmental factors, and lifestyle variables.
3. Taking care of the class imbalance: The dataset employed in this study has a problem with class imbalance, with most of the observations falling into the negative class. Future research can look into ways to solve this problem, for using cost-sensitive learning algorithms or over- or under sampling techniques.
4. Incorporating real-time diagnosis: To provide a real-time diagnosis of hypothyroidism, the models created in this study can be incorporated into clinical decision support systems. Future research can examine the viability, efficacy, and effects of such systems on patient outcomes.
5. Investigating transfer learning: Transfer learning is a method that uses pre-trained models as a jumping off point for a new assignment. Future research can look towards adapting previously trained models for other medical diseases to diagnose hypothyroidism via transfer learning.

Overall, the machine learning analysis on the hypothyroid dataset conducted in this paper has laid a solid groundwork for further investigation. By investigating these topics, scientists might enhance the precision and speed of hypothyroidism diagnosis, which could ultimately result in better patient outcomes.

References:

- https://indjst.org/articles/machine-learning-techniques-for-thyroid-disease-diagnosis-a-review
- http://papersim.com/wp-content/uploads/Neural_Network_Thyroid_Disease_2015.pdf
- https://iopscience.iop.org/article/10.1088/1742-6596/1963/1/012140/pdf
- https://www.mdpi.com/2072-6694/14/16/3914
- https://www.researchgate.net/publication/341534298_Thyroid_Disease_Prediction_Using_Machine_Learning_Approaches
- https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9405591/
- https://www.researchgate.net/publication/306007162_Machine_Learning_Techniques_for_Thyroid_Disease_Diagnosis_-_A_Review