2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Prediction of Thyroid Classes Using Feature Selection of AEHOA Based CNN Model for Healthy Lifestyle

Rachappa Jopate*[1], Piyush Kumar Pareek[2], DivyaJyothi M. G[1], Ariam Saleh Zuwayid Juma Al Hasani[1]

[1]Department of Information Technology, University of Technology and Applied Sciences, Al Mussanah, Oman.
[2]Department of AI ML, NITTE Meenakshi Institute of Technology, Bangalore, India.
*Corresponding Author.
ICCDA2023: International Conference on Computing and Data Analytics 2023.

Received 28/12/2023, Revised 26/04/2024, Accepted 28/04/2024, Published 25/05/2024

Abstract

People with underactive thyroids frequently endure severe symptoms. Correct classification and machine learning substantially improve thyroid disease diagnosis. This precise classification will impact the timely delivery of care to the patients. Although diagnostic techniques exist, they frequently seek binary categorization, use insufficiently big datasets, and lack confirmation of their conclusions. The focus of current approaches is on model optimisation, whereas feature engineering is neglected. This research presents the Adaptive Elephant Herd Optimisation Algorithm (AEHOA) model for selecting optimal attributes in order to circumvent these limitations. At first, employ a method called the Synthetic Minority Over-sampling Technique (SMOTE) to even out the data. Finally, the parameters of the AEHOA model are fed into a Convolutional Neural Network (CNN) to categorise data and enhance prediction. The accuracy of classification predictions was also increased by tweaking the dataset. Both datasets were put through a categorization process for a more precise comparison of results.

Keywords: Adaptive Elephant Herd Optimization Algorithm, Convolutional Neural Network, Hyperthyroidism Imbalanced data, Machine Learning, Synthetic Minority Over-sampling Technique.

Introduction

The healthcare sector is making use of computational biology developments by accumulating patient data for the sake of illness prediction. There are a sum of available tools for early illness diagnosis[1]. Intelligent requests, i.e., the evidence of medical knowledge, are not readily available to collect the necessary sets of data for illness analysis[2]. Recently, however, a method known as Machine Learning (ML) optimisation has emerged, making significant contributions to the prediction and resolution of non-linear and complicated problems. Maximum weight is given to characteristics from many datasets that may be selected in any illness detection strategy and readily categorise in healthy people[3]. Instead, a healthy individual may get unneeded therapy if they were incorrectly labelled as having a disease. Therefore, it is of utmost importance to accurately forecast any disorders in addition to thyroid[4].

The thyroid gland is an endocrine gland located in the neck. It develops in the human neck below the Adam's apple and helps the thyroid secrete

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

hormones, which in turn regulates protein synthesis and metabolic rate[5]. Thyroid hormones help regulate several aspects of metabolic rate[6], including heart rate and energy expenditure. Thyroid hormones are secreted by the thyroid gland and help regulate metabolism. Thyroid glands actively release the thyroid hormones triiodothyronine (T3) and levothyroxine (T4) [7]. These hormones play a significant role in the manufacturing process and in the supervisory framework as a whole by controlling core body temperature. The thyroid glands ordinarily generate two active hormones, T4 (also known as thyroxin) and T3 [8]. These hormones have crucial roles in the body's energy storage, temperature control, communication, and protein management. Deficiencies in T3 and T4 (two thyroid hormones) are associated with iodine deficiency, which is considered a fundamental component of the thyroid glands[9]. Both too little and too much of the thyroid hormones have a role in the development of hypothyroidism and hyperthyroidism, respectively. Several pathologies can lead to either underactive or overactive thyroids. Different medications have different uses[10]. Thyroid surgery puts patients at risk for iodine insufficiency, enzyme deficiency, ionising radiation exposure, and continuing thyroid pain.

Related works

Different machine learning (ML) procedures, including a scaling technique, an oversampling strategy, and other feature selection techniques, have been developed by Sultana and Islam[14] to provide a useful framework for classifying TD. In addition, key TD risk variables were identified using this methodology. The dataset used in this study was obtained from a database maintained by the University of California, Irvine (UCI). After that, the preprocessing step saw the application of SMOTE to fix the uneven classes and the robust scaling method to normalise the data. It employed the Boruta, RFE, and LASSO methods to narrow down the characteristics that were most relevant to the problem at hand. It was used six different machine learning classifiers to train the model. A 5-fold CV was used to analyse the models. The algorithms were evaluated using a variety of performance indicators. Using the RF classifier, the system was able to produce 99% correct results. The suggested method would help doctors and patients alike categorise TD and get insight into its related risk factors.

To tackle these problems, Yu et al. [15] present PyMLViT, a new approach with two fundamental modules: a transformer model. In order to gather probable background material on smears, create a pyramid-shaped token extraction module. Multi-scale local aspects are extracted by the pyramid token structure, and global information is obtained by the vision transformer structure via the self-attention mechanism. In addition, using the standard multi-instance learning framework, build a multi-loss fusion module. To increase the variety of supervised information, it was used slide-level participate in training, and to do so using carefully planned bag and patch weight allocation algorithms. Extensive experimental findings on the real-world dataset demonstrate that PyMLViT outperforms common approaches for identifying thyroid cancer in

Ultrasound imaging technology has allowed for the accurate diagnosis of a variety of thyroid disorders[11,12]. Clinical diagnosis using ultrasound involves collecting the patient's Thyroid Ultrasound Standard Deep Learning (DL) is typically broken down into two stages: first, the methodology, is used to train the sample images and the extract network is used to either classify or identify the sample images[13].

In order to categorise data for thyroid illness, researchers prioritised the feature selection strategy for eliminating superfluous characteristics. At first, the SMOTE method is used to rectify the unbalanced data. Next, the AEHOA model chooses the pertinent characteristics, and finally, CNN is employed to classify the data. The remaining sections of the paper are as shadows: This paper is organised as follows: Section 2 provides an impression of the literature; Section 3 details the suggested model; Section 4 discusses the experimental analysis; and Section 5 draws conclusions and suggests directions for further research.

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

cytological smears while still having a competitive number of parameters.

Patients with thyroid illness may be classified into several groups using a multiclass classification model proposed by Alnaggar et al. [16], which relies on XGBoost optimisation. The primary contributions are (i) the proposal of a Multiclass-Classification for the goal of diagnosing three distinct thyroid disorders, and (ii) the improvement in the accuracy of feature selection for classification using the row dataset., and (iv) improve upon the results of previously conducted research. Thyroid illness data from the UCI is used to teach and evaluate XGBoost. In addition, constructed the model using hyperparameter optimisation to attain and compare the greatest possible accuracy score. The findings demonstrate that, when compared to the state-of-the-art models, the optimised XGBoost outperformed them by a significant margin (99% accuracy).

A unique transfer learning approach using a distant domain high-level feature fusion (DHFF) model is suggested by Tang et al. [17]. As a result, the model may acquire more relevant transfer information while avoiding unnecessary feature fusion thanks to a smaller distribution gap between the source and destination domains. Multiple studies using both datasets verify the DHFF. Based on the findings, DHFF's classification accuracy with auxiliary source domains may reach 88.92%, which is an improvement of up to 8% over prior transfer and remote transfer methods.

A crucial role in Thyroid illness detection is played by a framework design proposed by Sinha et al. [18] that uses LightGBM, Sequential Backward Selection (SBS), and a metaheuristic approach called Whale Optimisation (WO). The primary purpose of this study is to deliver a method that is both extremely precise and logical for identifying human thyroid problem. Despite the impressive outcomes of the several methods used to thyroid data sets, our literature review shows that the data actually utilised for illness identification is redundant, unpredictable, and lacking feature values. Detecting thyroid abnormalities and starting appropriate treatment as soon as possible is the goal of the proposed work, which would involve developing an expert advising system based on the Opti-LightGBM architecture. The suggested Opti-LightGBM model beats many state-of-the-art comparable models and achieves an accuracy of 99.75% on the Thyroid dataset.

An optimised convolutional neural network model is provided by Srivastava and Kumar [19] for the detection of thyroid nodules through a sum of deep learning strategies, including the geometry group-16. This study takes into account a total of 295 available and 654 gathered thyroid ultrasound datasets. Data from 1475 publicly available and 3270 privately acquired thyroid ultrasound datasets are used to test the suggested model. Through experimentation, it was found the optimal values for the learning rate and drop out factor to improve the models' overall efficiency. Experiment-I results for the projected model on the public dataset show an accuracy of 93.75 percent, sensitivity of 94.6 percent, specificity of 92.5 percent, and f-measure of 94.0 percent; experiment-II results for the collected dataset show an accuracy of 96.89 percent, sensitivity of 97.80 percent, specificity of 94.7 percent, and f-measure of 97.2 percent. The suggested model outperforms state-of-the-art models on in terms of accuracy, sensitivity, specificity, and f-measure by margins of (4.57%, 7.84%), (5.06%, 8.24%), (4.43%, 6.63%), and (4.66%, 7.83%), respectively.

Proposed System

This research paper tourist attractions the importance of accurate classification of heart disease.

A. Dataset brief description

The experiment makes use of the thyroid dataset from the UCI ML repository. From the Garavan Institute in Sydney, Australia comes this contribution from Ross Quinlan [20]. There are a total of 3772 entries in the dataset, 2800 of which training and 972 for testing. The dataset contains a total of 29 characteristics, the last of which is a prognosis of the illness. There are 7 numerical characteristics and 22 category ones. The terms "age", "sex", "pregnancy", "goitre", "tumour", and "hypopituitarism" are just 11 of the many clinical variables that might be assessed. The collection also includes six test findings labelled

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

"TSH," "T3," "TT4", "T4U," "free thyroxine index," and "TBG." The model's process is depicted in Fig. 1.

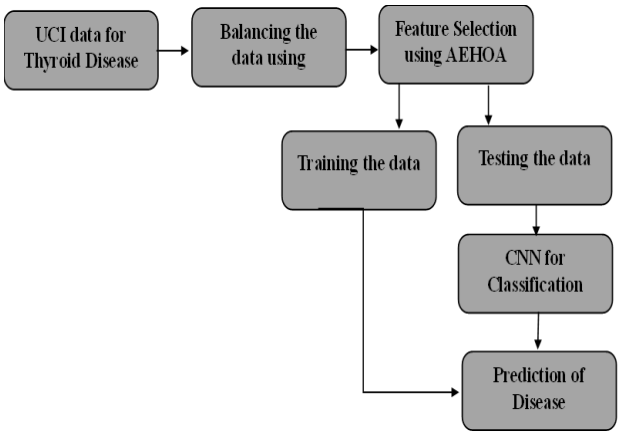

Figure 1. Working flow of the proposed model.

The thyroid dataset includes numerous missing values and is severely skewed towards one of four classifications. Out of the total 3772 cases in the dataset, 3481 (92.3%) are in the negative category, meaning they are normal. Another 194 (5.1%) are in the compensated-hypothyroid category, meaning they are hyperthyroid. Another 95 (2.5%) are in the primary-hypothyroid category, meaning they are hypothyroid. Finally, 2 (0.05%) are in the secondary-hypothyroid category, meaning they are underactive thyroids.

B. Preprocessing

The column means are substituted for missing data in the dataset. In a highly unbalanced dataset, there is a large discrepancy between the amounts of data in each category. An unbalanced dataset is useless. It is far more expensive to incorrectly diagnose a cancer patient as healthy than it is to incorrectly diagnose a healthy patient as having cancer in medical analysis. A false negative error might result in the loss of a human life, making it significantly costlier than a false positive error.

The (SMOTE) technique uses oversampling in these necessary cases[21-25], creating artificial records or tuples that are comparable to the provided dataset.

The collection of positive tuples that have been provided. The SMOTE technique is used to equalise the sum of datasets used in this investigation. In order to make sure there are an equal sum of instances, the number of thyroid cases is inflated. Table 1 displays the number of classes and the distribution of those classes in both the original dataset and the dataset generated after using SMOTE.

Table 1. Data Balancing the Thyroid Dataset with SMOTE Procedure

Dataset	No. of classes	Total records	Negative (0- No disease)	Positive (1, 2, 3)
Actual Dataset	4	3772	3481	291
After SMOTE	4	6795	3481	3314

SMOTE was used to collect a total of 6795 records, 3314 (48.77%) of which were negative; 760 were classified as primary_hypothyroid, 1584 as secondary_hypothyroid, and 970 as compensated_hypothyroid. As a result, it approximates a ratio of 51:49, making for a dataset; the same ratio is also used to separate the training and test sets.

C. Feature Selection using AEHOA

AEHO comprises the subsequent expectations:

- The populace of clans. Every single clan comprises an exact sum of elephants.
- The elephants usually alone.
- Each clan is directed by its eldest matriarch.

In a herd of elephants, the matriarchs accept the best option while the males' positioning reveals the worst. There are j different elephant families. Under Em's watchful eye, the members of clan c always make their next step based on the generation's greatest fitness quotient. The Eq. 1 describes this operation.

$$P_{new}, E_m, u = P_{E_m,u} + a(P_{bset,E_m} - P_{E_m,u}) \times Rd \quad \text{.....1}$$

Here, $P_{bset,E_m,u}$ characterizes the novel site of u in c, $P_{E_m,u}$ characterizes the old position, P_{bset,E_m} designates the finest solution of E_m, $\alpha \in [0,1]$ is the procedure's limit that determines the influence of the matriarch, and Rd is a accidental sum in the later phases of the procedure. The finest elephant's site in clan P_{bset,E_m} is efficient by applying Eq. 2.

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

$P_{new,E_{m,}d} = \beta \times P_{center,E_m}$ …….2

Here, $\beta \in [0,1]$ symbolizes the second the effect of P_{center,E_m}, as given by Eq. 3.

$P_{center,E_m}d = \frac{1}{uE\_} \times \sum_{j=1}^{u,E_m} b_{E_m,j,d}$ …….3

Here, 1 ≤ d ≤ D u_(E_m) represents the total sum of elephants in clan c, whereas D represents the total dimension of space.

The male elephants that wander away from their herd are put to use in scientific models. Some of the least suited elephants in each clan c are reassigned to new jobs, as stated in Eq. 4.

$P_{worst,E_m} = P(Pmin_{max} \times rand)_{min}$ …….4

where: $P_{\min}$ - search space, $P_{\max}$- search space and $rand \in [0,1]$- A accidental sum drawn from the unchanging distribution.

Crossover and mutation procedures are achieved during the evaluation of elephant locations to further optimise the system. Specifically, a crossover with two points is used. In this technique, pick two locations on each set of paternal chromosomes. The genes between these two locations are swapped out during reproduction, resulting in the offspring's chromosomes. The evaluation of these locations is given by Eq. 5 and Eq. 6.

$x1 = \frac{|P_{new,E_m}|}{3}$ …….5

$x2 = x1 + \frac{|P_{new,E_m}|}{2}$ …….6

The mutation is carried out by replacing several genes on each chromosome with new ones. Genes that have been randomly produced with no chromosomal repeat are the ones that have been switched. To improve the fitness value, this procedure is repeated.

The pseudocode of the AEHO procedure is given as underneath:

Input: *Weight values* ***Output***: *Optimized weight values*

Begin

Initialization

Set generation counter $t = 1$,

Set Maximum Generation M_G

Population $X_i = \{X_1, X_2 \dots \dots X_n\}$

While $t < M_G$ *do*

Sort all the elephants according to their fitness

for *all clans* E_m *in the population do*

for *all elephants* j *in the clan* E_m *do*

Update $P_{E_{m,}n} =$ *and generate* $P_{new,E_{m,}n}$ *by using,* $a_{new,E_{m,}n} = a_{E_{m,}n} + a(a_{bestE_m - a_{E_{m,}v}}) \times r$

if $P_{E_{m,}n} = P_{new,E_{m,}}$ *then*

Update $P_{E_{m,}n}$ *and generate* $P_{new,E_{m,}n}$ *by using* $P_{new,E_{m,}} = \beta \times P_{center,E_m}$

end if

end for

end for

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

for *all clans* $E_{m,}$ *in the population do*

Perform crossover and mutation

Replace the worst elephant in clan $E_{m,}$ *by using* P_{worst,E_m}
$= P(Pmin_{max} \times rand)_{min} Pworst; Em\ D\ P\ .Pminmax \times rand/min$

end for

Evaluate population by the newly updated positions. $t = t + 1$

end while

return *the best solution among all population*

End

Table 2 presents the selected features that are used in this work by using AEHOA model,

Table 2. Demonstrations the Features Limited in the Dataset

Clarification	Value Category	Attribute Name
1,2,3……..,	number	id
1,10,20,50,……,	number	age
1=m,0=f	1,0	gender
1=yes,0=no	1,0	query_thyroxine
1=yes,0=no	1,0	on_antithyroid_medication
1=yes,0=no	1,0	sick
1=yes,0=no	1,0	pregnant
1=yes,0=no	1,0	query_hyperthyroid
1=yes,0=no	1,0	TSH measured
Numeric value	Analysis ratio	TSH
1=yes,0=no	1,0	T3 measured
Numeric rate	Analysis ratio	T3
1=yes,0=no	1,0	T4 measured
Numeric value	Analysis ratio	T4
0=normal,1=hypothyroid, 2=hyperthyroid	0,1,2	category

Classification

Starting with the pre-processed data, the larger one is assembled and then the dataset is resampled to fit into the defined size by using AEHOA model. Pre-processed data is divided into training (80%) and testing data images (20%). In order to create a trained perfect, the ideal is first applied to training data.

In the projected effort, the model's Score are tested using a variety of standard metrics. By analysing selected data, the suggested network can categorise thyroid-infected patients. Our design calls for the use of a deep neural network with multiple layers and filtering capabilities. Each of the nine LeakyReLU layers, along with the other five dense layers, one flattening layer, and four dropout layers, is separated by three convolutional layers. A LeakyReLU operation is performed on each convolutional layer before it is sent to the layer. In this architecture, LeakyReLu is used as an activation function. The suggested design uses 2 x 2 and 3 x 3 kernels. Increasing the number of filters to 32, 64, and 256 is done on a regular basis. In the end, the activation layer generates the outputs in this model. Number of Epoch is 100, Number of Batch is 25, Optimizer used is Adamax and Model is used as “sequential”

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

Table 3. Proposed CNN Layer Defined

Layer (type)	Layer (type)	Output Shape	Output Shape	Parameter	Parameter
conv2d layer	conv2d_24	(255,255,32)	(255,255,32)	416	416
max_pooling2d	max_pooling2d_24	(127,127,32)	(127,127,32)	0	
conv2d_1	conv2d_25	(125,125,64)	(125,125,64)	18496	18496
batch_normalization	batch_normalization_14	(255,255,32)	(255,255,32)	128	128
leaky_re_lu	leaky_re_lu_32	(255,255,32)	(255,255,32)	0	0
conv2d_2	conv2d_26	(39,39,256)	(39,39,256)	147712	147712
batch_normalization layer _2	batch_normalization_16	(39,39,256)	(39,39,256)	1024	1024
batch_normalization layer _1	batch_normalization_15	(125,125,64)	(125,125,64)	256	256
leaky_re_lu_1	leaky_re_lu_33	(125,125,64)	(125,125,64)	0	0
max_pooling2d_1	max_pooling2d_25	(41,41,64)	(41,41,64)	0	
leaky_re_lu_2	leaky_re_lu_34	(39,39,256)	(39,256)	0	0
max_pooling2d_2	max_pooling2d_26	(19,19,256)	(19,256)	0	0
Fatten layer	fatten_8	(92416)	(92416)	0	0
dense_1	dense_41	(128)	(128)	32896	32896
leaky_re_lu_4	leaky_re_lu_36	(128)	(128)	0	0
Dropout layer _1	dropout_33	(128)	(128)	0	0
Dense layer _2	dense_42	(64)	(64)	8256	8256
leaky_re_lu_5	leaky_re_lu_37	(64)	(64)	0	0
dropout_2	dropout_34	(64)	(64)	0	0
Dense layer	dense_40	(256)	(256)	23658752	23658752
leaky_re_lu_3	leaky_re_lu_35	(256)	(256)	0	0
Dropout layer	dropout_32	(256)	(256)	0	0
Dense layer _3	dense_43	(32)	(32)	2080	2080
leaky_re_lu_6	leaky_re_lu_38	(32)	(32)	0	0
batch_normalization layer _3	batch_normalization_17	(32)	(32)	128	128
Dropout layer _3	dropout_35	(32)	(32)	0	0
Dense layer _4	dense_44	(3)	(2)	66	99

Results and Discussion

All tests were conducted on a 3.6 GHz Intel Core i9- with 16 GB of RAM and an NVIDIA GTX 2080Ti graphics dispensation unit. In addition, it was duse the accelerated computing resources of CUDA10.1 and cuDNN7.6.5 in the deep learning framework Pytorch.

A. Evaluation Metrics

Accuracy: "ratio of the observation of exactly predicted to the whole observations". This is exposed in Eq. 7.

$$T^{accuracy} = \frac{(Tr^p + Tr^n)}{Tr^p + Tr^n + Fa^p + Fa^n} \quad \text{.....7}$$

Sensitivity: "the number of true positives, which are recognized exactly".

$$Se = \frac{Tr^p}{Tr^p + Fa^n} \quad \text{.....8}$$

Specificity: "the number of true negatives, which are determined precisely".

$$Sp = \frac{Tr^n}{Fa^n} \quad \text{.....9}$$

Precision: "the ratio of positive observations that are predicted exactly to the total number of observations that are positively predicted".

$$Pr = \frac{Tr^p}{Tr^p + Tr^p} \quad \text{.....10}$$

FPR: "the ratio of count of false positive predictions to the entire count of negative predictions".

$$FPR = \frac{Fa^p}{Fa^p + Tr^n} \quad \text{.....11}$$

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

FNR: "the proportion of positives which yield negative test outcomes with the test".

$$FNR = \frac{Fa^n}{Tr^n+Tr^p} \quad \text{.....12}$$

NPV: "probability that subjects with a negative screening test truly don't have the disease".

$$NPV = \frac{Fa^n}{Fa^n+Tr^p} \quad \text{.....13}$$

FDR: "the number of false positives in all of the rejected hypotheses".

$$FDR = \frac{Fa^p}{Fa^p+Tr^p} \quad \text{.....14}$$

F1 score: It is distinct as the "harmonic mean between precision and recall. It is used as a statistical measure to rate performance".

$$F1score = \frac{Se.Pr}{Pr+Se} \quad \text{.....15}$$

MCC: It is a "correlation coefficient computed by four values".

$$MCC = \frac{Tr^p \times Tr^n - Fa^p \times Fa^n}{\sqrt{(Tr^p+Fa^p)(Tr^p+Fa^n)(Tr^n+Fa^p)(Tr^p+Fa^n)}} \quad \text{.....16}$$

Table 4. Analysis of Various Techniques for Thyroid Disease Prediction

Metrics	MLP	DBN	LR	AE	Proposed
FPR	0.10526	0.80645	0.18421	0.89655	0.078947
MCC	0.68803	0.72464	0.55436	0.77612	0.76555
FDR	0.11765	0.14211	0.2	0.2105	0.085715
Sensitivity	0.78947	0.75	0.73684	0.80263	0.84212
Specificity	0.89474	0.65789	0.81579	0.68421	0.92105
FNR	0.21053	0.15789	0.26316	0.078947	0.15789
NPV	0.89474	0.34211	0.81579	0.31579	0.92105
Precision	0.88235	0.84211	0.8	0.92105	0.91429
F1-Score	0.83333	0.79355	0.76712	0.70345	0.87672
Accuracy	0.84211	0.86241	0.77642	0.85174	0.88258

In above, Table 4 represents the Analysis of Various Practises for Disease Prediction. In the analysis of different models and different metrics are used as FDR metrics of LR model achieved as 0.2 and then MLP model attained as 0.11765 and then the AE model attained as 0.2105 and further DBN model attained as 0.14211 and finally the proposed model attained as the metrics value as 0.085715 correspondingly. After the F1-Score metrics of LR model achieved as 0.76712 and then MLP model attained as 0.83333 and then the AE model attained as 0.70345 and further DBN model attained as 0.79355 model attained as the metrics value as 0.87672 correspondingly. After the accuracy metrics of LR model achieved as 0.77642 and then MLP model attained as 0.84211 and then the AE model attained as 0.85174 and further DBN model attained as 0.86241 model attained as the metrics value as 0.88258 correspondingly. After the Sensitivity metrics of LR model achieved as 0.73684 and then MLP model attained as 0.78947 and then the AE model attained as 0.80263 and further DBN model attained as 0.84212 correspondingly. After the specificity metrics of LR model achieved as 0.81579 and then MLP model attained as 0.89474 and then the AE model attained as 0.68421 and further DBN model attained as 0.65789 0.92105 correspondingly. After the FNR metrics of LR model achieved as 0.26316 and then MLP model attained as 0.21053 and then the AE model attained as 0.15789 and further DBN model attained as model attained as the metrics value as 0.15789 correspondingly. After the NPV metrics of LR model achieved as 0.81579 and then the AE model attained as 0.31579 model attained as the metrics value as 0.34211 model attained as the metrics value as 0.92105 correspondingly. After the precision metrics of LR model achieved as 0.8 and then MLP model attained as 0.88235 and then the AE model attained as 0.92105 and further DBN model attained as 0.84211 model attained as the metrics value as 0.91429 correspondingly. After the FPR metrics of LR model

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

achieved as 0.18421 and then MLP model attained as 0.10526 and then the AE model attained as 0.89655 and further DBN model attained as 0.80645 model attained as the metrics value as 0.078947 correspondingly. After the MCC metrics of LR model achieved as 0.55436 and then MLP model attained as 0.68803 and then the AE model attained as 0.77612 and further DBN model attained as 0.72464 model attained as the metrics value as 0.76555 correspondingly. Fig. 2 shows the graphical representations of the various models discussed. The comparative analysis of the different models used are shown in Fig. 3. Also, the error analysis of the proposed model is represented in Fig. 4, showing the proposed model to be more efficient.

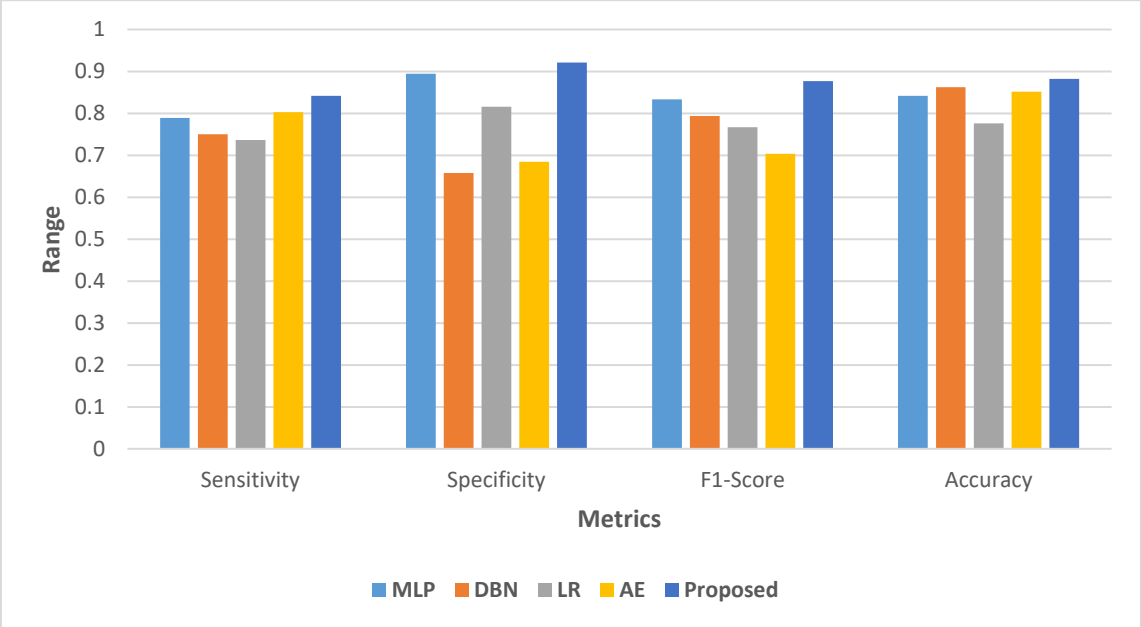

Figure 2. Graphical Representation of various models.

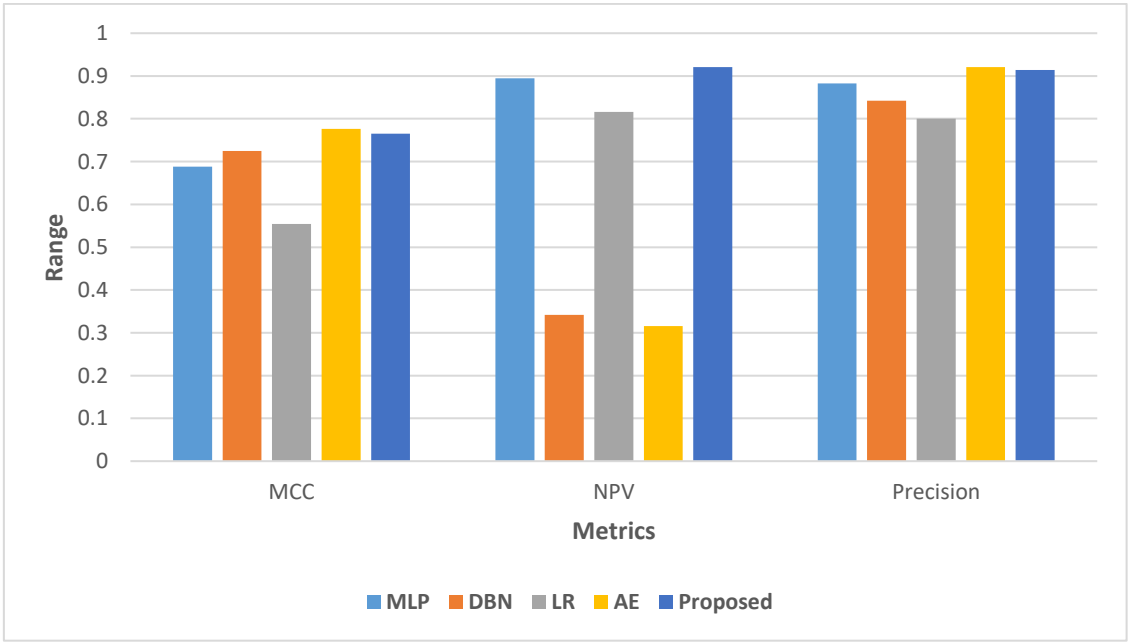

Figure 3. Comparative Analysis of different models.

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

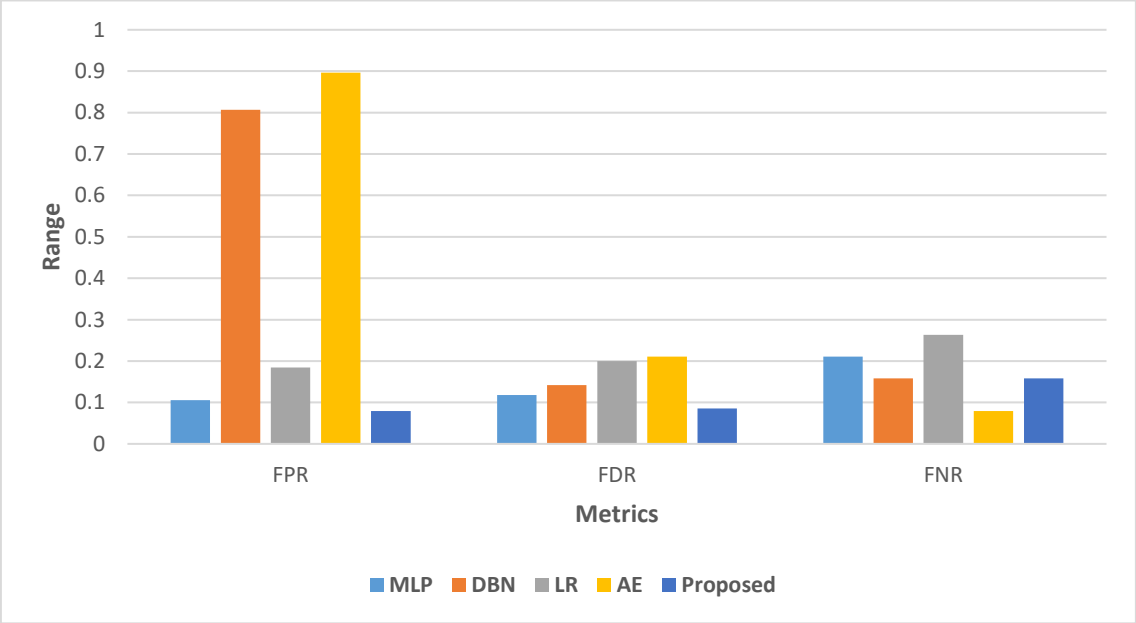

Figure 4. Error analysis of proposed model.

Conclusion

Thyroid illness identification is becoming increasingly urgent. Neither the models nor the dataset used to evaluate them have been thoroughly verified. The research presented here addresses these drawbacks by suggesting a method that makes use of feature selection with a deep learning model. Future trends in the medical use of AI include the autonomous measurement of thyroid parameters. In this study, a novel DL procedure was used to predict the identification and categorization of thyroid illness. After SMOTE is used to normalise the data in preparation for the AEHOA model's optimal feature selection, the data is ready for analysis. The classification accuracy of both the data is then improved through the application of CNN. The results show that the AEHOA model is most effective when combined with characteristics picked using a convolutional neural network (CNN). Due to their condensed computational complexity, CNNs are promising for use in the prediction of thyroid illness. These findings are supported by a 10-fold cross-validation test. When practises, the suggested method shows considerable improvement in performance. In the future, plan to expand our work to a classification job, where it will generate a model can properly forecast the pixel tags for a thyroid condition using only a little quantity of training statistics.

Authors' Declaration

- Conflicts of Interest: None.
- We hereby confirm that all the Figures and Tables in the manuscript are ours. Furthermore, any Figures and images, that are not ours, have been included with the necessary permission for re-publication, which is attached to the manuscript.
- No animal studies are present in the manuscript.
- No human studies are present in the manuscript.
- Ethical Clearance: The project was approved by the local ethical committee at University of Technology and Applied Sciences, Oman.

Authors' Contribution Statement

R. J., P. K. P., D. M.G and A. S. Z. J. contributed to the design and implementation of the research, to the analysis of the results and to the writing of the manuscript.

References

1. Kumar N, Nandihal P, B MR, Pareek PK, T N, R SS. A Novel Machine Learning-Based Artificial Voice Box. In: 2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE) 2022; Bangalore, India. IEEE. 2022; p. 1-7. https://doi.org/10.1109/ICATIECE56365.2022.10046967.
2. Nandihal P, Shetty VS, Guha T, Pareek PK. Glioma Detection using Improved Artificial Neural Network in MRI Images. In: 2022 IEEE 2nd Mysore Sub Section International Conference (MysuruCon); 2022; Mysuru, India. IEEE. 2022;p. 1-9. https://doi.org/10.1109/MysuruCon55714.2022.9972712.
3. Subbalakshmi C, Pareek PK, Narayana MV. A Gravitational Search Algorithm Study on Text Summarization Using NLP. Artificial Intelligence and Data Science. ICAIDS 2021. Communications in Computer and Information Science. Cham. 2022; 1673. https://doi.org/10.1007/978-3-031-21385-4_13.
4. Luna-Guevara JJ, Arenas-Hernandez MMP, Martínez de la Peña C, Silva JL, Luna-Guevara ML. The role of pathogenic E. coli in fresh vegetables: behavior, contamination factors, and preventive measures. Int J Microbiol. 2019; Article ID 2894328. https://doi.org/10.1155/2019/2894328
5. Ula M, Pratama A, Asbar Y, Fuadi W, Fajri R, Hardi R. A New Model of The Student Attendance Monitoring System Using RFID Technology. J Phys Conf Ser. 2021; 1807(1): 012026. https://doi.org/10.1088/1742-6596/1807/1/012026
6. Ramteke B, Dongre S. IoT Based Smart Automated Poultry Farm Management System. In: 2022 10th International Conference on Emerging Trends in Engineering and Technology-Signal and Information Processing (ICETET-SIP-22); 2022; pp. 1-4. https://doi.org/10.1109/ICETET-SIP-2254415.2022.9791653
7. Doaa Mohey El-Din Mohamed Hussein. A survey on sentiment analysis challenges. J King Saud Univ - Eng Sci. 2018; 30(4): 330-338. ISSN 1018-3639. https://doi.org/10.1016/j.jksues.2016.04.002
8. Hora SK, Poongodan R, Perez de Prado R, Wozniak M, Divakarachari PB. Long short-term memory network-based metaheuristic for effective electric energy consumption prediction. Appl. Sci. 2021; 11(23): 11263. https://doi.org/10.3390/app112311263
9. Dey S, Ye Q, Sampalli S. A machine learning based intrusion detection scheme for data fusion in mobile clouds involving heterogeneous client networks. Inf Fusion. 2019; 49: 205-215. https://doi.org/10.1016/j.inffus.2019.01.002
10. Karthik R, Radhakrishnan M, Rajalakshmi R, Raymann J, Manjunath R, Kwadiki K. Delineation of ischemic lesion from brain MRI using attention gated fully convolutional network. Biomed Eng. Lett. 2021; 11: 3-13. https://doi.org/10.1007/s13534-020-00178-1
11. Mallik A, Khetarpal A, Kumar S. ConRec: malware classification using convolutional recurrence. J Comput. Virol. Hacking Tech. 2022; 1-17. https://doi.org/ 10.1007/s11416-022-00416-3
12. Han Y, Liu M, Jing W. Aspect-level drug reviews sentiment analysis based on double BiGRU and knowledge transfer. IEEE Access. 2020; 8: 21314-21325. https://doi.org/10.1109/ACCESS.2020.2969473
13. Alqahtani A, Alqahtani N, Alsulami AA, Ojo S, Shukla PK, Pandit SV, et al. Classifying electroencephalogram signals using an innovative and effective machine learning method based on chaotic elephant herding optimum. Expert Syst. 2023: e13383. https://doi.org/10.1111/exsy.13383.
14. Garcia-Perez A, Cegarra-Navarro JG, Sallos MP, Martinez-Caro E, Chinnaswamy A. Resilience in healthcare systems: Cyber security and digital transformation. Technovation. 2023; 121: 102583. ISSN 0166-4972. https://doi.org/10.1016/j.technovation.2022.102583.
15. Sulaiman R, Schetinin V, Sant P. Review of Machine Learning Approach on Credit Card Fraud Detection. Hum-Cent Intell Syst. 2022; 2: 55–68. https://doi.org/10.1007/s44230-022-00004-0.
16. Alnaggar, M., Handosa, M., Medhat, T., Z. Rashad, M. Thyroid Disease Multi-class Classification based on Optimized Gradient Boosting Model. Egypt. J Artif Intell. 2023; 2(1): 1-14. https://doi.org/10.21608/ejai.2023.205554.1008
17. Tang F, Ding J, Wang L, Ning C. A Novel Distant Domain Transfer Learning Framework for Thyroid Image Classification. Neural Process Lett. 2023; 55(3): 2175-2191. https://doi.org/ 10.1007/s11063-022-10940-4
18. Sinha BB, Ahsan M, Dhanalakshmi R. Light GBM empowered by whale optimization for thyroid disease detection. Int J Inf Technol. 2023: 1-10. https://doi.org/10.1007/s41870-023-01261-3.
19. Srivastava R, Kumar P. Optimizing CNN based model for thyroid nodule classification using data augmentation, segmentation and boundary detection techniques. Multimed Tools Appl. 2023: 1-36. https://doi.org/10.1007/s11042-023-15068-8.

2024, 21(5 Special Issue): 1786-1797
https://doi.org/10.21123/bsj.2024.10547
P-ISSN: 2078-8665 - E-ISSN: 2411-7986
Baghdad Science Journal

20. Huang SF, Cheng CH. A safe-region imputation method for handling medical data with missing values. Symmetry. 2020; 12(11): 1792. https://doi.org/10.3390/sym12111792.
21. Amirruddin AD, Muharam FM, Ismail MH, Tan NP, Ismail MF. Synthetic Minority Over-sampling TEchnique (SMOTE) and Logistic Model Tree (LMT)- sufficiency levels of oil palm (Elaeis guineensis) using spectroradiometers and unmanned aerial vehicles. Comput Electron Agric. 2022; 193: 106646. https://doi.org/10.1016/j.compag.2021.106646.
22. Sen S, Jopate R, Kerur SS, Manjunatha LH, Ahmad A, Jothiprakash G. Nanocomposites for Energy Storage. Materials for Sustainable Energy Storage at the Nanoscale. 1st ed. CRC Press. 2023; p. 331-336. https://doi.org/10.1201/9781003355755.
23. Falhi A, Luaibi N, Alsaedi A. Hypothyroidism and AMH in Iraqi Patients with Chronic Kidney Disease. Baghdad Sci J. 2021; 18(Suppl. 1): 695-699. https://doi.org/ 10.21123/bsj.2021.18.1(Suppl.).0695.
24. Alnedawe SM, Aljobouri HK. A New Model Design for Combating COVID -19 Pandemic Based on SVM and CNN Approaches. Baghdad Sci J. 2023 Aug. 1 [cited 2024 Jan. 14]; 20(4): 1402. https://orcid.org/0000-0003-1361-2607.
25. Bao Y, Yang S. Two Novel SMOTE Methods for Solving Imbalanced Classification Problems. IEEE Access. 2023; 11: 5816-5823. https://doi.org/10.1109/ACCESS.2023.3236794.

التنبؤ بفئات الغدة الدرقية باستخدام اختيار الميزات لنموذج CNN القائم على AEHOA لأسلوب حياة صحي

راشابا جوبات[1]، بيوش كومار باريك[2]، ديفيا جيوثي إم جي[1]، أريام صالح زويد جمعة الحسني[1]

[1]قسم تقنية المعلومات، الجامعة التقنية والعلوم التطبيقية، المصنعة، عمان.
[2]قسم الذكاء الاصطناعي والتعلم الآلي، معهد NITTE ميناكشي للتكنولوجيا، بنغالور، الهند.

الخلاصة

كثيرًا ما يعاني الأشخاص الذين يعانون من قصور الغدة الدرقية من أعراض حادة. يؤدي التصنيف الصحيح والتعلم الآلي إلى تحسين تشخيص أمراض الغدة الدرقية بشكل كبير. سيؤثر هذا التصنيف الدقيق على تقديم الرعاية للمرضى في الوقت المناسب. على الرغم من وجود تقنيات التشخيص، فإنها تسعى في كثير من الأحيان إلى التصنيف الثنائي، وتستخدم مجموعات بيانات كبيرة غير كافية، وتفتقر إلى تأكيد استنتاجاتها. تركز الأساليب الحالية على تحسين النموذج، في حين يتم إهمال هندسة الميزات. يقدم هذا البحث نموذج خوارزمية تحسين قطيع الفيل التكيفي AEHOA لاختيار السمات المثالية من أجل التحايل على هذه القيود. في البداية، استخدم طريقة تسمى تقنية الإفراط في أخذ العينات للأقلية الاصطناعية SMOTE لتسوية البيانات. وأخيرًا، يتم إدخال معلمات نموذج AEHOA في الشبكة العصبية التلافيفية CNN لتصنيف البيانات وتعزيز التنبؤ. تمت أيضًا زيادة دقة تنبؤات التصنيف عن طريق تعديل مجموعة البيانات. تم إخضاع مجموعتي البيانات لعملية تصنيف لإجراء مقارنة أكثر دقة للنتائج.

الكلمات المفتاحية: خوارزمية تحسين قطيع الفيل التكيفية، الشبكة العصبية التلافيفية، فرط نشاط الغدة الدرقية، بيانات غير متوازنة، تقنية الإفراط في أخذ العينات للأقليات الاصطناعية.