Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Improving Medical Diagnosis Using Missing Data Treatment Techniques: A Case Study on Thyroid Data

Sajjad Basim Abdulyasser*

Wasit,52001,Iraq.

**Email: sajjadbassim95@gmail.com*

ARTICLE INFO

Article history:
Received: 2 /2/2025
Rrevised form: 13 /2/2025
Accepted : 2 /3/2025
Available online: 30 /3/2025

Keywords:

Maximum expectation(EM), Multiple chained input(MICE), Missing data, Classification model, Thyroid diseases, ML modeling , Random forests.

ABSTRACT

In this paper, we focus on some advanced input techniques that deal with missing data from thyroid medical datasets such as maximum expectation (EM), multiple sequential input (MICE), etc. It is demonstrated that EM and MICE had a significant positive impact on prediction accuracy compared to their traditional counterparts (i.e. direct and ensemble inputs). Moreover, EM significantly enhanced predictions derived from random forests in a manner consistent with previous findings of additive predictive power for clinical variables only, confirming the promise of EM (and MICE) to provide a significant fundamental improvement in modeling in complex medical models, in the case of a comprehensive analysis of phenotypes similar to heart failure. In conclusion, the results of the present study indicate that advanced input methods increase the accuracy of diagnosis and therapeutic predictions in patients with thyroid diseases.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.11974

1. Introduction

the process of correctly managing missing values, is important research challenge, given that missing data can undermine the statistical validity of the data. Expectation maximization and multiple imputation with chained equations are two of the most commonly used methods to tackle this type of problem, and provide insight into how to model missing values as well as helping to fill in the gaps. The expectation maximization algorithm is iterative, estimating the parameters of a statistical model with the available data, generating estimates, and using them to fill missing data, while multiple imputation using chained equations uses a series of regression models to generate multiple, plausible values for a missing data point. [16]. This research paper is primarily aimed at delivering a rigorous review of these two techniques as well as comparing the benefits, disadvantages and effectiveness in a range of case studies. The crux of the study is to compare the performance of SVM, DT, KNN and RF models, before and after applying Expectation Maximization and Multiple Imputation techniques. While Expectation Maximization (EM) and Multiple Inference with Chained Equations (MICE) were applied to our data, we did perform both direct inference and Clustered inference on all four stated algorithms. The direct imputation method replaces missing

*Corresponding author : Sajjad Basim Abdulyasser

Email addresses: : sajjadbassim95@gmail.com

Communicated by 'sub etitor'

values with the expected values based on the machine learning models. whereas the ensemble imputation method leverages the underlying data structure to cluster similar observations and imputed the missing values based on that clustering. Next, the paper goes on to describe the performance of the models, discussing what was done and explaining the metrics used to evaluate the models: accuracy, precision, recall and F1 score. The findings suggest that the integration of machine learning algorithms with sophisticated imputation methods can lead to substantial enhancements in the predictive accuracy across the board, particularly when encountering missing data scenarios.

2. Related Works

This paper relates the diagnostic accuracy of data to the performance of machine learning algorithms with health conditions of thyroid disorders. With a high-quality dataset obtained from the KEEL repository, machine learning models have been shown to achieve significant results in diagnostic accuracy [2]. The study presented analyses and comparisons of machine learning models in predicting the probability of thyroid cancer recurrence, and achieved prediction accuracy results of 93% [3].

SVM machine learning and logistic regression was used to analyze the thyroid dataset and discussed an approach. A comparison between these two algorithms was conducted in terms of precision, recall, F-measure, ROC, and RMS error. Ultimately, the best classification method turned out to be logistic regression [4].

In this paper, they show the MCSVM approach for diagnosing and classifying hypothyroidism, as well as the results indicate the One-Against-All technique with polynomial kernels provide the highest accuracy among all techniques. A maximum accuracy of 96.9% was obtained in data tests which proved MCSVM to be a very successful and accurate method of classifying hypothyroid cases[5].

Shrila Dash et al. propose an enhanced classification model for hypothyroidism risk prediction: they develop a data mining model for hypothyroidism classification and prediction. The focus is on deploying various data mining techniques to enhance the process of classification of the intended input data using the dataset available at the University of California repository. In summary, the proposed model combining preprocessing and data augmentations before the classification process is able to increase the prediction accuracy, recall,and the accuracy of hypothyroid diagnosis. Our specialization in hypothyroidism is of particular interest because patients with hypothyroidism often present with atypical symptoms, making diagnosis more difficult and potentially delaying treatment and improving patient outcomes in the management of hypothyroidism [6].

Based on data from the UCI Machine Learning Repository, R.P. Ram Kumar et al. The study titled “Thyroid Disease Classification using Machine Learning Algorithms”(Tashfeen, W. 2023) explores how machine learning techniques can improve thyroid disease detection. Focusing on hyperthyroidism and hypothyroidism, the paper analyzes the performance of several algorithms for classification, e.g., the Decision Trees, Support Vector Machines, and Neural Networks, in the diagnosis of thyroid disorders [7].

Khalid Salman and Ampullas Sonuç presented a study which aims to find out the different between hyperthyroidism, hypothyroidism and normal thyroid functions through the machine learning models. Popular algorithms are used in the research to analyze thyroid disease data in Iraqi patients, such as support vector machines, random forests, decision trees, naive Bayes algorithm, logistic regression, K-nearest neighbors, multilayer perceptron's, and linear discriminant analysis. This comprehensive strategy is designed to enhance the precision of thyroid disease classification to enable optimal diagnosis and treatment [8].

Esra’a Alshdaifat talked about the impact of preprocessing techniques on classification algorithms' performance. She is discussed proper preprocessing technique that would help the classification model in order to be properly configured to achieve better accuracy and can also help in reducing computation if used properly, these steps would help when the features are numerical. Normalization transformed the data like a scalar making between the range of 0-1 (min-max transformation) and how many standard deviations from the mean (z-value normalization) proved to impact positively the sucess rate of a classifier as in order to reduce the noise of the model [9].

The research is focused on the efficient and reliable extraction of knowledge from the construction of operational data. Preparing Data for Energy Management: Reason for Data Preprocessing In the paper, the authors present a comprehensive overview of a variety of data preprocessing approaches, which seek to enhance the quality of data for more robust and efficient analysis. The methods include missing value imputation, outlier detection, data reduction,

expansion, transformation, and segmentation. The paper also presents advanced data science and its techniques, data augmentation, transfer learning, and semi-supervised learning, which are fast and useful ways to tackle real issues of data analytics [10].

3. Methodology

3.1. Dataset description

The Garavan Institute Thyroid Diseases dataset, accessible from the UCI Machine Learning Repository, encompasses a comprehensive collection of data aimed at facilitating the study and prediction of various thyroid conditions. The dataset comprises 3772 instances, categorized into several thyroid health statuses based on clinical measurements and patient information.

The dataset features 29 attributes plus a class label that categorizes each instance into different types of thyroid conditions such as negative (no disease), compensated hypothyroid, primary hypothyroid, and secondary hypothyroid. The attributes can be broadly classified into two types: binary (Boolean) and continuous, providing a diverse range of data points for analysis.

Column	Description
Class	Class label indicating the type of thyroid disease: negative, compensated hypothyroid, primary hypothyroid, or secondary hypothyroid.
TSH_measured	Indicates whether TSH (Thyroid Stimulating Hormone) level was measured (Yes/No).
TSH_reading	The measured level of TSH, essential for thyroid function assessment.
T3_measured	Indicates whether T3 (Triiodothyronine) level was measured (Yes/No).
T3_reading	The measured level of T3, important for metabolic assessment.
T4_measured	Indicates whether T4 (Thyroxine) level was measured (Yes/No).
T4_reading	The measured level of T4, crucial for metabolic function and thyroid health.
FTI_measured	Indicates whether FTI (Free Thyroxine Index) was measured (Yes/No).
FTI_reading	The measured level of FTI, indicative of the amount of free, active thyroid hormone.
Age	The age of the patient, as age can influence thyroid function.
Sex	The sex of the patient, given the prevalence of thyroid disorders differs by gender.
On_thyroxine	Indicates if the patient is currently taking thyroxine medication.
Query_on	Query regarding the patient's use of thyroxine, used to adjust diagnosis or treatment.

_thyroxine	
On_antithyroid _medication	Indicates if the patient is taking anti-thyroid medications which can affect thyroid hormone levels.
Thyroid_surgery	Indicates if the patient has undergone thyroid surgery, which can drastically affect thyroid function.
Query_hypothyroid	Query if there is a suspicion or diagnosis of hypothyroidism based on symptoms or test results.
Query_hyperthyroid	Query if there is a suspicion or diagnosis of hyperthyroidism based on symptoms or test results.
Pregnant	Indicates if the patient is pregnant; pregnancy can complicate the diagnosis of thyroid disorders.
Sick	Indicates if the patient is currently experiencing any illness, which might affect thyroid function tests.
Tumor	Indicates if the patient has a tumor, which could potentially affect the thyroid gland.
Lithium	Indicates if the patient is taking lithium, a drug known to impact thyroid function.
Goitre	Indicates the presence of goitre, an enlargement of the thyroid gland.
Psych	Indicates if the patient has a psychological condition, which might correlate with thyroid health.
TSH	Specific reading of Thyroid Stimulating Hormone level in the blood.
T3	Specific reading of Triiodothyronine level in the blood.
TT4	Specific reading of Total Thyroxine (TT4) level in the blood.
T4U	Specific reading of Thyroxine Utilization rate.
FTI	Specific reading of Free Thyroxine Index.
TBG	Specific reading of Thyroxine-binding globulin level, not commonly measured.

3.2. Data Pre-Processing

3.1. **Data Cleaning:** In this task information is cleaned by removing information from redundant information and resolving information irregularities. For this reason, data quality is improved because data is insufficiency [19]. We did the preprocessing of the dataset, out of total 3772 samples in this dataset, 61 duplicate records contain in the dataset. The rest 3711 patient records are used to classify a person as with a thyroid disorder (coded 1) or without a thyroid disorder (coded 0). If the patient has a thyroid disorder, the value of the feature is one, and if not, zero.

3.2. **Missing Values:** They see a limited subset of features even though there might be a hundred other relevant features. If your variables are sometimes observable and sometimes not, then yes, you can use the cases in which this variable is observable to predict information about the classes in which the variable is not present. This is often called dealing with missing data ML algorithms can, therefore, discover patterns and relationships from the observed data using the instances where the variable is observable. In such cases, the learned patterns are then used to predict the values of the variable in instances that have either missing data or are unobserved. In this paper, we will discuss two types of algorithms that we can use to improve the data and compensate for missing values that may not be observed, which are:

1 **Expectation-Maximization (EM):**

The EM algorithm is a statistical method used for finding maximum likelihood estimates of parameters in probabilistic models, where the model depends on unobserved latent variables. Instead, we use the widely applicable algorithm that has been used in Bayesian analysis of spatial data, estimation and imputation of missing data in longitudinal models, and computing generalized closed-form maximum likelihood estimators by randomly-picking unobserved latent points. The Expectation-Maximization algorithm has one of its primary applications in dealing with missing data. When data is not completely observed, the algorithm can be used to estimate the missing values and simultaneously update the model parameters. This approach is particularly useful in situations where traditional methods, such as listwise deletion or mean imputation, may lead to biased estimates [17].

The main idea behind EM algorithm is to estimate the missing data by utilizing the available observed data of the dataset and then using that data to update the parameters of the data. There are many more details that we will have to go through to understand EM algorithm.

- **Init-Step:** First, a range of initial values of the parameters are taken. The system is assumed to get a measured dynamics coming from a certain model with only a subset of the measured data
- **E-Step:** during E-step we use the observed data to estimate or predict the missing or incomplete data. It is just used to add the variables.
  1. Compute Stand: P (all else) to incorporate the previous knowledge about what we knew about the data and Five Parameter Estimates.
  2. The missing patterns are imputed (or the equivalent of it) predicted from the existing parameters.

  Estimate log-likelihood of the data given the parameters and estimated hidden data (E step).

- **M-Step:** In this step, we update the values of the parameters using the complete information that we obtained in the previous step (E step). Mostly, itis used to review the theory.
  1 Optimize the E-step expected complete data log-likelihood to update the model parameters.
  2 Usually this means optimization problems, where parameter values that maximize the log-likelihood are sought

- **Convergence -Step:** The convergence stage checks if the values tend to converge. If they are, the process finishes if not follow it back to points 2 and 3, which is the "E" step and the "M" step one more time until we converge.
  1 Monitor convergence through examining log-likelihood or other parameter variation across iterations
  2 If change is less than a minimum threshold, K, exit and assume converged.
  3 Instead, go back to the E-step and repeat until convergence.

2 **Multiple imputation by chained equations**

Multiple Imputation by Chained Equations, also known as Fully Conditional Specification, is a robust and versatile technique for handling missing data in research studies. Unlike single imputation methods, which tend to underestimate the uncertainty of the imputed values, multiple imputation generates multiple plausible estimates for each missing value, allowing for better quantification of the uncertainty associated with the imputation process [23]. The MICE approach involves an iterative procedure, where the missing values are imputed based on the observed data and previously imputed values MICE Works :

1. Dummy Data: For this step, you can replace all the N/A inputs on the table with simple placeholders, e.g., mean/median of each variable (mode for categorical type).

2. Reset Missing Values: The next step is to take one variable at a time for imputation, and take those placeholder values back to missing.
3. Step 3: Create Regression Model: Build a regression model using the independent variables (excluding the dependent variable from Step 2) to predict the variable identified in Step 2. It is a model that will utilize all (or some of rather) the other variables in the dataset as predictors of the missing data point values.
4. Impute Missing Values: Employ the regression model to estimate and populate the missing values for the said variable. Now we have both observed and newly imputed values for this variable.
5. Iterate for Every Variable: Perform the same 2-4 procedure for every variable in the dataset containing unknowns.

3.3. Data Partitioning:

Splitting randomly separates the dataset into two groups, which the first will only use as training data and another test data. One creates and trains a model with training data, which is then validated against testing data[20]. Following this study, the researcher chooses to split the exercise with a 70-30 ratio, where 70 % goes to coaching and 30 % to a validation.

3.4. Data Transformation:

This is the process of converting information or data from a specific format to another format. It usually happens when a source format needs to be converted into the required format for a certain goal. This includes aggregation, standardization or normalization and smoothing [18]. Normalization This happens when we apply rescaling to a real number attribute to the range of [0,1].

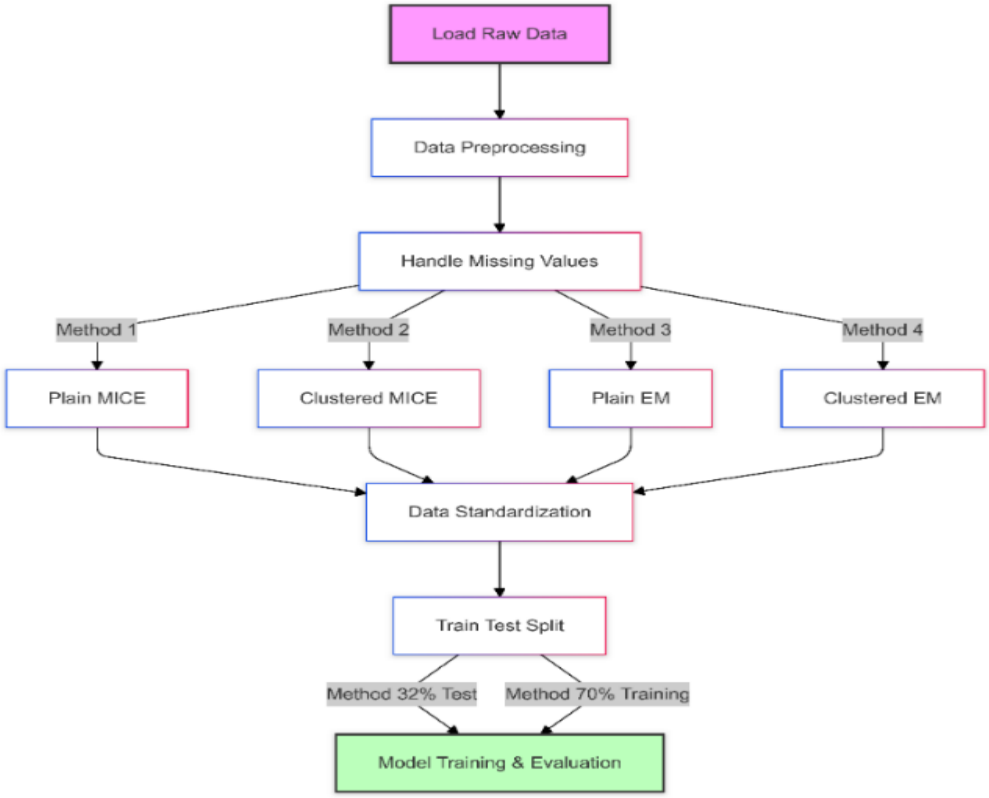

FIGURE 1. *for*

preprocessing

3.3. Classification Analysis

Machine learning has a plethora of applications, but one of the most fundamental tasks is through classification; this comes into play in everything from image recognition to natural language processing. The objective of this work was to perform a comparative evaluation of the SVM, DT, RF and KNN classifiers applied on the real-world datasets.

- **Random Forest**

  We examine Random Forest, a widely used ensemble learning technique for addressing classification and regression issues. Its applications are numerous: pattern recognition, feature extraction, picture segmentation, function approximation, and data mining. [25] The Random Forest technique generates several Decision Trees using diverse random subsets of training data and amalgamates the prediction outputs from each tree to ascertain the final prediction. The random forest algorithms.

  1) Obtain n bootstrap samples from the original dataset.

  2) An unpruned classification or regression tree is developed for each bootstrap sample, with the change that at each node, instead of choosing the best split from all predictors, a random subset of predictors is picked from which the ideal split is determined [26].

- **Decision Trees**

  I keep saying popular Decision Trees Machine learning algorithm you can use it for classification and regression problems These models work by following a sequence of tests or decisions to predict the output class or value for a given input [27].In Brief — Decision Trees splits the data recursively according to the most informative feature of the dataset at hand making decision results a tree structure. Early decision tree work includes the classic "Classification and Regression Trees" book by Bierman et al., which introduced the basic principles of decision trees and their uses [28].

- **K-Nearest Neighbors**

  K-Nearest Neighbors (KNN) is one of the most commonly used supervised machine learning algorithm. It is a versatile algorithm that can be applied to both classification and regression problems [29]. The basic idea for KNN is that it finds the k closest training examples to a test example, and classifies or predicts the target variable based on the majority the class (for classification problems) or average of the target variable (for regression problems) of the k nearest neighbors. [29]

- **Support Vector Machines**

  Support Vector Machines, as a powerful and versatile class of supervised learning algorithms, have now become one of the most popular supervised algorithms in the machine learning field as they can effectively solve many classifications and regression task.
  Support Vector Machines support vector machines work by trying to find hyperplane which best separates the classes by maximizing the margin or gap. [30][31] It does this by transforming the input data into a high-dimensional feature space where the data is linearly separable.
  The SVM has one main advantage which it can work with a very complicated, mess up dataset/vector. SVMs can also capture complex patterns and interactions within the data by using alternative Kernel functions like the polynomial kernel or the radial basis function kernel [30][32].

- **LR algorithm**

  Logistic regression is a relatively simple yet powerful technique that is commonly used in machine learning and statistical datamining for binary classification problems. Linear regression is inappropriate for these analyses as it is used when the dependent variable is continuous, while in logistic regression we will be using a dichotomous dependent variable, one that can take only two values such as “yes” or “no,” “success” or “failure,” or “admitted” and “rejected".One of the main strengths of logistic regression is that it can be used to model the probability of a positive event, conditional on a set of predictor variables. searches to the unstructured "type" of text [39]. As the most used machine learning algorithm, it remains a powerful option for many applications, including predicting customer churn, detecting fraudulent transactions, classifying medical diagnosis, or voting outcome predictions [38].

- **Gradient Boosting Machines**
  Due to their impressive classification potential, the gradient boosting machines have received much attention many papers on this topic [40].
  Gradient Boosting Machines is an ensemble learning technique that establishes a prediction model by centering on the strengths of multiple weak learners, often decision trees [41]. Gradient boosting machines are based on the idea that each subsequent model can improve upon the previous step by minimizing the loss function, which is often a measure of the difference between predictions and the true outcome [40].
- **LDA**
  Linear Discriminant Analysis (LDA) is a popular supervised learning algorithm applied in classifications problems, especially for high dimensional datasets. LDA redirects the linear combination of the input features to maximize the separation between different classes so more effective classification can be achieved [42].
  LDA is well suited for multiclass classification (assignment of an input to one out of multiple classes possible) [42].
- **AdaBoost**
  The AdaBoost classification algorithm, also known as Adaptive Boosting, is a powerful machine learning technique that combines multiple weak classifiers to create a strong and accurate classifier. AdaBoost works by iteratively training weak classifiers, such as decision trees, and then combining them to form a strong ensemble [43].

3.4. Hyperparameter optimization

hyperparameter tuning is a crucial step in the machine learning pipeline, as it can have a significant impact on the performance of a model [33]. Manually selecting hyperparameters can be a time-consuming and inefficient process, especially as the number of hyperparameters increases. Optuna, a powerful open-source hyperparameter optimization framework, offers a solution to this problem by automating the hyperparameter tuning process. Optuna's Bayesian optimization approach allows for efficient exploration of the hyperparameter space, providing a systematic way to identify the optimal hyperparameter configurations for a given model [34] [35]. This process involves modeling the objective function (the model's performance) as a sample from a Gaussian process and then using this model to guide the selection of the next hyperparameter configuration to evaluate [35]. Existing research on hyperparameter optimization frameworks has highlighted the advantages of Bayesian optimization over traditional methods like grid search and random search [36] [37]. Bayesian optimization has been shown to obtain better results in fewer evaluations compared to these non-adaptive approaches, due to its ability to reason about the quality of experiments before they are run [37].

4. Result

Algorithms Performance Analysis with Data Missing Values Handling Techniques

Ensemble learning methods outperform their competitors in experimental results. AdaBoost yielded outstanding performance (99.89%) on Clustered MICE, and GBM exhibited stable high performance (>99.58%) in all situations.

Random Forest produced solid results, especially Clustered EM (99.89%) confirming the prediction that clustering and missing data imputation will work better together.

The following table shows the test results obtained in all algorithms

Algorithm	Plain EM	Plain MICE	Clustered EM	Clustered MICE
Random Forest	99.47%	98.62%	99.89%	99.68%
SVM	97.35%	96.92%	97.24%	97.03%
K-NN	97.35%	95.12%	94.91%	95.12%
Decision Trees	92.79%	96.82%	89.50%	98.83%
AdaBoost	99.58%	99.89%	99.79%	99.89%
GBM	99.79%	99.68%	99.58%	99.79%
LDA	94.3%	95.1%	94.9%	94.7%
LR	97.56%	97.88%	97.56%	97.56%

Table 1 Result of algorithms

5. Conclusion

The study of advanced input techniques for handling missing data in thyroid medical analysis has proven its effectiveness in decision making processes, where a number of key points have been highlighted that highlight the great importance of these techniques in enhancing medical accuracy diagnosis and efficiency of prediction models.

In this paper, we have demonstrated that advanced input using Maximum Expectation (EM) and Multiple Chained Input (MICE) techniques provide good and noticeable improvements in accuracy compared to direct and ensemble input.

These results reflect the great ability of these techniques to deal with the complexities and relationships in medical data, as they contributed to extracting accurate and valuable information from missing data, which is of great importance in developing data-driven medical models.

References

[1] Ashok, L. and Sivanandam, S. (2017) Diagnosis of Thyroid Disorder Using Infrared Thermography. International Conference of Electronics, Communication and Aerospace Technonlogy (ICECA), Coimbatore, 20-22 April 2017, 37-41.

[2] S. Balasubramanian, "Clinical Relevance of ML Predictions using Health Datasets," University of Victoria Repository, 2024.

[3] S. Alam, M. A. Hider, and A. Al Mukaddim, "Machine Learning Models for Predicting Thyroid Cancer Recurrence," Journal of Medical Informatics, 2024.

[4] Y. F. Wang, Comparison Study of Radiomics and Deep-Learning Based on Methods for Thyroid Nodules Classification Using Ultrasound Images, 8, IEEE Access, (2020)

[5] F. F. Chamasemani and Y. P. Singh, “Multi-class Support Vector Machine (SVM) Classifiers -- An Application in Hypothyroid Detection and Classification,” Theories and Applications, pp. 351–356, Sep. 2011, doi: 10.1109/bic-ta.2011.51.

[6] S. Dash, M. N. Das, B. K. Mishra, School of Computer Engineering, KIIT University, Bhubaneswar-751024,Odisha, India, and Department of IT, C.V. Raman College of Engineering, Bhubaneswar-752054,Odisha, India, “Implementation of an optimized classification model for prediction of hypothyroid disease risks,” journal-article, 2021.

[7] R. P. R. Kumar, M. S. Lakshmi, B. S. Ashwak, K. Rajeshwari, and S. M. Zaid, “Thyroid Disease Classification using Machine Learning Algorithms,” E3S Web of Conferences, vol. 391, p. 01141, Jan. 2023, doi: 10.1051/e3sconf/202339101141.

[8] K. Salman and E. Sonuç, “Thyroid disease classification using machine learning algorithms,” Journal of Physics Conference Series, vol. 1963, no. 1, p. 012140, Jul. 2021, doi: 10.1088/1742-6596/1963/1/012140.

[9] E. Alshdaifat, D. Alshdaifat, A. Alsarhan, F. Hussein, and S. M. F. S. El-Salhi, “The effect of preprocessing techniques, applied to numeric features, on classification algorithms’ performance,” Data, vol. 6, no. 2, p. 11, Jan. 2021, doi: 10.3390/data6020011.

[10] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, “A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building Operational data,” Frontiers in Energy Research, vol. 9, Mar. 2021, doi: 10.3389/fenrg.2021.652801.

[12] B. Srilatha, H. B. A, and D. Soumya, “Epidemiology and Treatment for Thyroid Cancer,” Jan. 01, 2011, OMICS Publishing Group. doi: 10.4172/1948-5956.s17-011.

[13] K. A. Kelly and D. Brady, “Emergencies in Thyroid Function,” Sep. 11, 2009, Elsevier BV. doi: 10.1016/j.jen.2009.04.012.

[14] R. Shames, “Diagnostic Challenges and Treatment Options for Thyroid Conditions,” Feb. 01, 2012, Mary Ann Liebert, Inc. doi: 10.1089/act.2012.18109.

[15] R. Brown, J. A. de Souza, and E. E. W. Cohen, “Thyroid Cancer: Burden of Illness and Management of Disease,” Jan. 01, 2011, Ivyspring International Publisher. doi: 10.7150/jca.2.193.

[16] M. Azur, E. A. Stuart, C. Frangakis, and P. J. Leaf, “Multiple imputation by chained equations: what is it and how does it work?,” Feb. 24, 2011, Wiley. doi: 10.1002/mpr.329.

[17] V. Tadayon, “Bayesian Analysis of Censored Spatial Data Based on a Non-Gaussian Model,” Mar. 01, 2017. doi: 10.18869/acadpub.jsri.13.2.155.

[18] D. Shah, S. Patel, and S. K. Bharti, "Heart Disease Prediction using Machine Learning Techniques," SN Computer Science, vol. 1, no. 6, 2020, doi: 10.1007/s42979-020-00365-y.

[19] C. B. Gokulnath and S. P. Shantharajah, "An optimized feature selection based on genetic approach and support vector machine for heart disease," Cluster Computing, vol. 22, no. S6, pp. 14777-14787, 2018, doi: 10.1007/s10586-018-2416-4.

[20] A. A. Ali, "Stroke Prediction using Distributed Machine Learning Based on Apache Spark," Stroke, vol. 28, no. 15, pp. 89-97, 2019. [Online]. Available:https://www.researchgate.net/profile/Nahla-Omran2/publication/338458550_Stroke_Prediction_using_Distributed_Machine_Learning_Based_on_Apache_Spark/links/5e1619404585159aa4be6a2e/Stroke-Prediction-using-Distributed-Machine-Learning-Based-on-Apache-Spark.pdf.

[21] P. Koturwar, S. Girase, and D. Mukhopadhyay, “A Survey of Classification Techniques in the Area of Big Data,” Jan. 01, 2015, Cornell University. doi: 10.48550/arxiv.1503.07477.

[22] G. G. Towell, “Using Unlabeled Data for Supervised Learning,” Nov. 27, 1995. Accessed: Dec. 2024. [Online]. Available: https://papers.nips.cc/paper/1097-using-unlabeled-data-for-supervised-learning.pdf

[23] L. Gondara and K. Wang, “MIDA: Multiple Imputation using Denoising Autoencoders,” Jan. 01, 2017, Cornell University. doi: 10.48550/arxiv.1705.02737.

[24] A. Choudhury and M. R. Kosorok, “Missing Data Imputation for Classification Problems,” Jan. 01, 2020, Cornell University. doi: 10.48550/arxiv.2002.10709.

[25] H. Wang, “Pattern Classification with Random Decision Forest,” Aug. 01, 2012. doi: 10.1109/icicee.2012.42.

[26] A. Liaw and M. C. Wiener, “Classification and Regression by randomForest,” Jan. 01, 2007. Accessed: Dec. 2024. [Online]. Available: http://cogns.northwestern.edu/cbmg/LiawAndWiener2002.pdf

[27] C. Kingsford and S. L. Salzberg, “What are decision trees?,” Nature Biotechnology, vol. 26, no. 9. Nature Portfolio, p. 1011, Sep. 01, 2008. doi: 10.1038/nbt0908-1011.

[28] A. Criminisi, Decision Forests: A Unified Framework for Classification, Regression, Density Estimation, Manifold Learning and Semi-Supervised Learning. 2011. doi: 10.1561/9781601985415.

[29] L. Kozma, “k Nearest Neighbors algorithm (kNN),” Jan. 01, 2008. Accessed: Jan. 2025. [Online]. Available: http://www.cis.hut.fi/Opinnot/T-61.6020/2008/knn.pdf

[30] S. N. Wright and T. Marwala, “Artificial Intelligence Techniques for Steam Generator Modelling,” Jan. 01, 2008, Cornell University. doi: 10.48550/arxiv.0811.1711.

[31] J. M. Moguerza and A. Muñoz, “Support Vector Machines with Applications,” Aug. 01, 2006, Institute of Mathematical Statistics. doi: 10.1214/088342306000000493.

[32] G. L. Prajapati and A. Patle, “On Performing Classification Using SVM with Radial Basis and Polynomial Kernel Functions,” Nov. 01, 2010. doi: 10.1109/icetet.2010.134.

[33] P. I. Frazier, “A Tutorial on Bayesian Optimization,” Jan. 01, 2018, Cornell University. doi: 10.48550/arxiv.1807.02811.

[34] J. Bergstra, R. Bardenet, Y. Bengio, and B. Kégl, “Algorithms for Hyper-Parameter Optimization,” Dec. 12, 2011, Centre National de la Recherche Scientifique. Accessed: Jan. 2025. [Online]. Available: https://hal.inria.fr/hal-00642998

[35] J. Snoek, H. Larochelle, and R. P. Adams, “Practical Bayesian Optimization of Machine Learning Algorithms,” Jan. 01, 2012, Cornell University. doi: 10.48550/arxiv.1206.2944.

[36] H. Wang, S. Gao, H. Zhang, W. Su, and M. Shen, “DP-HyPO: An Adaptive Private Hyperparameter Optimization Framework,” Jan. 01, 2023, Cornell University. doi: 10.48550/arxiv.2306.05734.

[37] M. Feurer and F. Hutter, “Hyperparameter Optimization,” in The Springer series on challenges in machine learning, Springer International Publishing, 2019, p. 3. doi: 10.1007/978-3-030-05318-5_1.

[38] David Hubbard,Benoît Rostykus,Yves Raimond,T. Jebara, “Beta Survival Models.” Oct. 2023. Accessed: Jan. 22, 2025. [Online]. Available: https://arxiv.org/pdf/1905.03818.pdf

[39] P. Ranganathan, C. S. Pramesh, and R. Aggarwal, “Common pitfalls in statistical analysis: Logistic regression,” Jan. 01, 2017, Medknow. doi: 10.4103/picr.picr_87_17.

[40] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, “A comparative analysis of gradient boosting algorithms,” Aug. 24, 2020, Springer Science+Business Media. doi: 10.1007/s10462-020-09896-5.

[41] N. Ponomareva et al., “TF Boosted Trees: A scalable TensorFlow based framework for gradient boosting,” Jan. 01, 2017, Cornell University. doi: 10.48550/arxiv.1710.11555.

[42] C. J. Huberty and R. M. Barton, “An Introduction to Discriminant Analysis,” Oct. 01, 1989, SAGE Publishing. doi: 10.1080/07481756.1989.12022925.

[43] D. W. Opitz and R. Maclin, “Popular Ensemble Methods: An Empirical Study,” Aug. 01, 1999, AI Access Foundation. doi: 10.1613/jair.614.