INVESTIGATING CLASSIFICATION METHOD FOR DATASET WITH MULTI FEATURES

Jeremy Ansellino Gunawan1\*, Patricia Citranegara Kusuma2, Kenjovan Nanggala3, Devi Fitrianah4

1,2,3,4Computer Science Department, BINUS Graduate Program

Master of Computer Science, Bina Nusantara University, Jakarta 11480 Indonesia

\*Email Penulis Koresponden: jeremy.gunawan001@binus.ac.id

|  |  |
| --- | --- |
| ***Abstract:***  In this study, we delve into the exploration of classification methods tailored for datasets that have specific characteristics. The characteristics of the dataset used are that it has many features, is not balanced, and has an output. These aspects represent actual data analysis challenges, where high-dimensional datasets can lead to dimensionality and complexity issues, and imbalanced classes may result in notable biases within the classification model. The datasets used are two datasets with the same characteristic the Indian Pima Dataset, and the Heart Disease Dataset UCI. The investigation encompasses seven prominent classification algorithms, comprising linear-based models (Support Vector Machine and Logistic Regression), tree-based models (Random Forest and Decision Tree, and XGBoost), probabilistic models (Gaussian Naive Bayes), and a deep learning model (Artificial Neural Networks). The primary objective is to meticulously assess the effectiveness of these methodologies in precisely categorizing instances within datasets characterized by a wide array of diverse feature types. In this study, we can conclude that tree based model especially Random Forest, are more suitable for the various multiple feature and imbalance datasets. The advantage lies in the ensemble approach, which overcomes class imbalance by constructing many random trees, strengthening minority classes. The Random Forest also determines relevant features in multi-feature datasets, and is tolerant of outliers, reducing negative impact. Its ability to adapt and generalize well to new data is also a key factor.  *This is an open access article under the* [*CC BY-NC*](https://creativecommons.org/licenses/by-nc/4.0/) *license* | ***Kata Kunci****:*  *Classification;*  *Multi-Feature;*  *Imbalance Dataset*  ***Riwayat Artikel:***  Received Jun x, 20xx  Revised Nov x, 20xx  Accepted Dec x, 20xx  **DOI**:  10.22441/incomtech.v10i3.7777 |

**1. INTRODUCTION**

In the current digital era, the exponential growth of data has given rise to new challenges and opportunities across various fields, including data science and machine learning. Classification, as one of the primary branches of machine learning, aims to categorize data into predetermined classes [1]. The success of classification plays a crucial role in various applications such as pattern recognition, spam detection, sentiment analysis, and more [2]. Effective classification models are key to uncovering insights hidden within increasingly complex datasets [3].

In an endeavor to develop more accurate and reliable classification models, this research aims to compare the performance of several widely-used classification models. These models include Logistic Regression, Random Forest, Support Vector Machine (SVM), Extreme Gradient Boosting (XGBoost), Decision Tree, Artificial Neural Network (ANN), and Gaussian Naive Bayes. Utilizing this diverse set of models enables a comprehensive evaluation of their capabilities in addressing classification challenges within datasets exhibiting specific characteristics [4].

The dataset chosen for this study has been thoughtfully selected to encompass distinctive features: a high number of dimensions and an imbalanced class distribution. These aspects mirror real-world challenges in data analysis, where high-dimensional datasets can pose issues of dimensionality and complexity, and imbalanced classes can lead to significant biases in classification models [5].

Datasets characterized by multiple features and class imbalance present a distinctive set of challenges within the domain of machine learning and data analysis. These challenges encompass several critical aspects, including model bias, the misclassification of minority class instances, and the potential distortion of evaluation metrics.Class imbalance, prevalent in such datasets, can introduce a bias in machine learning models, inclining them towards favoring the majority class, thereby undermining their ability to make accurate predictions for minority class instances. The misclassification of minority class samples becomes particularly concerning, especially in scenarios where the minority class signifies critical or rare events, such as fraud detection in financial transactions or the diagnosis of rare diseases in medical applications. Furthermore, the imbalance inherent in these datasets can mislead the evaluation process, as standard metrics like accuracy may not adequately reflect the model's true performance. Hence, the challenges associated with multiple-feature datasets and class imbalance demand specialized techniques and strategies to address these issues effectively.

Addressing the intricate challenges presented by datasets characterized by both class imbalance and multiple features is a crucial aspect of machine learning and data analysis. To tackle these challenges, a multifaceted approach is required. Firstly, resampling techniques such as oversampling, undersampling, or employing algorithms like SMOTE can rebalance the class distribution. Secondly, modifying loss functions to assign greater weight to minority class misclassifications can enhance model performance. Ensemble methods such as Random Forests and Gradient Boosting provide a robust means of handling imbalanced data, while anomaly detection algorithms offer an alternative perspective on identifying rare instances. Thoughtful feature engineering, including dimensionality reduction methods, can further accentuate class differences. Selecting appropriate algorithms, adjusting decision thresholds, and using cross-validation techniques that preserve class distribution integrity are also vital. The augmentation of minority class data, when feasible, and the application of cost-sensitive learning approaches add to the toolkit. Lastly, domain expertise plays a pivotal role in guiding these strategies, from preprocessing to model selection, ensuring that the most relevant features are emphasized. Rigorous evaluation with imbalanced dataset-specific metrics, such as precision, recall, and accuracy, is essential to assess the success of these methods in achieving a balanced and accurate classification outcome. To tackle the imbalance and multi feature dataset the author using preprocessing and selecting seven classification model to see the most effective result.

The primary goal of this research is to compare and analyze the performance of the aforementioned seven classification models on datasets with these distinctive characteristics (imbalance and multi feature). By considering appropriate evaluation metrics such as accuracy, precision, and recall, this study aims to provide in-depth insights into which model is most effective in tackling challenges associated with high-dimensional datasets and imbalanced class distributions.

In previous research, Sofia Benbelkacem and Baghdad Atmani [6] conducted an investigation utilizing the Pima Indians Diabetes dataset, employing the Random Forest algorithm as the primary analytical tool. In their study, multiple Random Forest models were constructed, varying in the number of decision trees to determine the optimal forest size. The outcomes of these experiments demonstrated the efficacy of their approach based on Random Forest. In comparison to alternative machine learning techniques, their methodology exhibited superior efficiency, underlining the suitability of Random Forest as a robust choice for predictive modeling and analysis of the Pima Indians Diabetes dataset. Furthermore, in a separate study by Madhumita Pal and Smita Parija [7], the Random Forest algorithm was applied to classify heart disease patients using the Heart Disease Dataset UCI. Their findings revealed an impressive accuracy rate of 86.9% for heart disease prediction, with a sensitivity value of 90.6% and a specificity value of 82.7%, underscoring the utility of Random Forest in medical classification tasks. approach to address imbalanced datasets within the Pima Indian Diabetes dataset was undertaken by Victor Chang, Jozeene Bailey, Qianwen Ariel Xu, and Zhili Sun [8], Their research encompassed the application of Naïve Bayes, Random Forest, and J48 models for classification. The study discerned that the Naïve Bayes model demonstrated notable performance, particularly when coupled with a refined feature selection process for binary classification. In contrast, the Random Forest model exhibited enhanced performance when a broader feature set was employed, offering valuable insights into the nuances of addressing class imbalance in the dataset.

In other words, research on disease classification using datasets with multi-feature and unbalanced characteristics typically involves a multifaceted approach to address these challenges. Some of the techniques frequently employed include deep feature fusion models, which utilize feature partition and deep feature extraction [9], and methods like SMOTE and Nearmiss to balance unbalanced data [10]. Feature selection techniques are commonly applied to choose a subset of characteristics with both low inner similarity and high relevance to the target class. Additionally, supervised machine learning algorithms are often employed to predict the prolonged length of stay of hospitalized patients diagnosed with various chronic conditions [11]. For specific diseases like diabetes and heart disease, specialized methods have been proposed. For diabetes, a hybrid sampling approach combining SMOTE and ENN aims to enhance the performance of Support Vector Machine and Random Forest classification methods [12]. Similarly, in heart disease diagnosis, a hybrid method integrates feature selection into ensemble classifier model generation, contributing to improved disease classification outcomes [13]. These diverse strategies collectively contribute to the effective handling of multi-feature, unbalanced datasets in disease classification research.

**2. METHODOLOGY**

This study embarks on an in-depth exploration of classification methodologies tailored to datasets characterized by specific characteristics, including the presence of numerous features, and class imbalance. The investigation centers on a comparative analysis of seven classification model techniques, such as Support Vector Machine, Logistic Regression, XGBoost, Gaussian Naive Bayes, Random Forest, Artificial Neural Networks, and Decision Tree. The primary objective is to address the unique data characteristics by assessing the efficacy of these models through precision, recall, and accuracy evaluations, ultimately culminating in the identification of the most suitable model.

As outlined in Figure 1, the research flow is methodically structured. Initiated by data collection from online sources, the subsequent stage involves Exploratory data analysis to identify and analyze general patterns in the data [14], followed by meticulous data cleaning and preprocessing to ensure data quality and relevance. Subsequently, the prepared data undergoes modeling using a diverse set of classification models, including SVM, Logistic Regression, XGBoost, Gaussian Naive Bayes, Random Forest, ANN, and Decision Tree.

Post-modeling, the research shifts towards a comprehensive analysis of the outcomes, allowing for a comparative evaluation of each algorithm's performance to see which model best classifies the data with the same characteristic. This evaluation serves as the basis for identifying the most effective model. The ensuing step involves an in-depth discussion segment, wherein conclusions, findings, and recommendations are extracted and synthesized.

A diagram of data collection

Description automatically generated

Figure. 1. Research Flow

**2.1 Data Collection**

In this study, our investigation is carried out across three datasets sharing similar characteristics.The similar characteristics means the dataset have multiple features, and an imbalance dataset. Our primary objective is to determine the classification model that aligns most suitable with the distinct attributes of these datasets. Specifically, the chosen datasets exhibit common traits: an abundance of features, an imbalanced distribution, and a definitive output. The three datasets used are Pima Indian diabetes Dataset, Heart Disease UCI Dataset, and Indian Liver Patient Dataset.

**2.1.1 Pima indian diabetes dataset**

This dataset was collected as part of a medical study aimed at identifying factors associated with the risk of diabetes within the Indian Pima population. It has become a commonly used data source in diabetes classification research.

**Data Source:** This dataset was obtained from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) and is available on the UCI Machine Learning Repository.

**Data Attributes:** This dataset consists of a number of attributes that encompass both medical and demographic information. These attributes include age, number of pregnancies, plasma glucose concentration, blood pressure, skinfold thickness, serum insulin, body mass index (BMI), family history of diabetes, and the two-hour serum glucose measurement following an oral glucose tolerance test. The dataset contains a total of **768 samples**.

**2.1.2 Heart disease UCI dataset**

This dataset was collected as part of a medical study aimed at identifying factors associated with the risk of heart disease within a population. It has become a commonly used data source in heart disease classification research.

**Data Source:** This dataset was obtained from the UCI Machine Learning Repository.

**Data Attributes:** This dataset consists of a number of attributes that encompass both medical and demographic information. These attributes include age, gender, blood pressure, cholesterol levels, fasting blood sugar levels, electrocardiogram (ECG) results, treadmill test outcomes, and more. The dataset contains a total of **303 samples.**

**2.2 Exploratory Data Analysis**

Exploratory Data Analysis (EDA) is a crucial phase in data analysis, serving as the initial step to gain insights and understand the underlying patterns within a dataset. EDA involves a variety of techniques and visualizations to summarize, describe, and uncover relationships in the data [15]. Through measures like central tendency and dispersion, EDA provides an overview of the dataset's distribution and variability. Graphical representations such as histograms, box plots, and scatter plots help identify outliers, trends, and potential data issues. Correlation matrices and heatmaps reveal interdependencies among variables. EDA also includes handling missing values, checking for data integrity, and assessing the need for data transformations[16]. By employing EDA, analysts can formulate hypotheses, guide further analyses, and make informed decisions about data preprocessing and modeling strategies [17]. This process ultimately aids in extracting valuable insights and enhancing the quality of subsequent analyses.

The correlation matrix of every feature of Pima Indian Dataset, Indian Liver Patient Dataset, and Heart Disease UCI Dataset can be seen on Figure 2-3.

**2.2.1 Pima Indian diabetes dataset**

The Pima Indian Diabetes dataset, originally sourced from the National Institute of Diabetes and Digestive and Kidney Diseases, contains detailed information about 768 women of Pima Indian heritage, all at least 21 years of age. The primary focus of the dataset is to determine the presence or absence of diabetes, with 258 women testing positive and 500 testing negatives. The target variable, "Outcome," is binary, with 1 indicating the presence of diabetes and 0 indicating the absence.

This dataset comprises eight numeric explanatory variables. Firstly, "Pregnancies'' represents the number of pregnancies experienced by each woman. Secondly, "Glucose" indicates the two-hour plasma glucose concentration measured in mg/dl following an oral glucose tolerance test. Thirdly, "BloodPressure '' signifies diastolic blood pressure recorded in mmHg. Fourthly, "SkinThickness" reflects triceps skinfold thickness measured in millimeters. Next, "Insulin" represents the 2-hour serum insulin levels measured in mu U/ml. The sixth variable is "BMI," which denotes the Body Mass Index measured in kg/m^2. Additionally, there is "DiabetesPedigreeFunction," a function that estimates the likelihood of diabetes based on ancestral history. Finally, "Age" denotes the age of individuals in years. It's important to highlight that this dataset contains missing values in various features. Specifically, there are instances where the minimum values for the "Glucose," "BloodPressure," "SkinThickness," "Insulin," and "BMI" columns are recorded as 0. These values are not physically possible and are likely indicators of missing data. To address this issue, it is recommended to replace these 0 values with NaN (Not-a-Number) to properly identify and handle them as missing values in the analysis.

In our data analysis, we observe that while there is no strong correlation among the features, some relatively stronger correlations are evident. Specifically, we find that Glucose and Insulin exhibit a correlation coefficient of 0.59, indicating that higher insulin levels coincide with elevated blood glucose levels. Age and Pregnancies show a correlation coefficient of 0.54, suggesting that older women tend to have a higher number of pregnancies. Moreover, Glucose and Outcome display a correlation coefficient of 0.48, implying that elevated glucose levels are associated with an increased likelihood of diabetes. Additionally, a noteworthy correlation of 0.63 exists between Skin Fold Thickness and BMI, implying that greater skin fold thickness is linked to higher BMI, potentially indicating a connection to overweight or obesity. We also identify a negative correlation of -0.031 between Diabetes Pedigree Function and Pregnancies. Importantly, there is no significant multicollinearity concern, as none of the correlations among the features exceed the threshold of 0.7, which could complicate predictive modeling.

Grouping the data by the Outcome reveals several patterns. Diabetic women tend to have more pregnancies, higher glucose levels, elevated blood pressure, greater skin thickness, higher insulin levels, increased BMI, and a higher Diabetes Pedigree Function score compared to non-diabetic women. Both groups show BMI values well above the normal range (18.5 - 25), indicating a prevalence of obesity. Diabetic women are more likely to have a family history of diabetes, and they tend to have insulin levels above the normal range (16–166 muU/mL), whereas non-diabetic women have insulin levels within the normal range. Additionally, both groups exhibit glucose levels higher than the normal range (<= 100 mg/dL), suggesting that some non-diabetic women may be at risk of developing diabetes in the future, especially those with elevated insulin levels.

Visualizing the data through line plots confirms that diabetic women generally have higher mean values for each feature compared to non-diabetic women. Before addressing missing values, it's crucial to consider data distribution. Features like Blood Pressure (BP), Skin Thickness, and BMI exhibit a normal distribution, while others are skewed. Mean imputation is suitable for normally distributed features, while skewed ones benefit from median imputation. Given the substantial missing values in Skin Thickness and Insulin, KNN imputation is chosen to handle these features due to the potential for mean/median imputation to generate significant outliers. The Age distribution is highly skewed, indicating a predominantly young population. Furthermore, outliers are detected in all features, necessitating their treatment for enhanced accuracy.

Pair plots facilitate the visualization of variable relationships and distributions. Notably, a strong linear relationship is observed between Skin Thickness and BMI, alongside a high correlation between Insulin and Glucose levels, aligning with correlation matrix results. Kdeplots further illustrate the linear association between Glucose and Insulin, both displaying right-skewed distributions. Joint plots, employing 'Outcome' as a hue parameter, highlight that diabetic women tend to exhibit higher values across features, implying older age, greater obesity, more pregnancies, elevated blood pressure, increased glucose and insulin levels, and thicker skin folds. Elevated glucose levels consistently emerge as a significant diabetes indicator, regardless of other variables.

To gauge feature importance in diabetes prediction, an Extra Tree Regressor is utilized, revealing Glucose as the most impactful, followed by BMI, Age, and Insulin levels. Principal Component Analysis (PCA) is harnessed for dimensionality reduction, transforming correlated variables into uncorrelated principal components while reducing dataset dimensionality. The explained variation per principal component is provided for further insight.

**2.2.2 Heart disease UCI dataset**

Cardiovascular diseases (CVDs), commonly known as heart diseases, rank as the primary global cause of death, claiming the lives of around 17.9 million people annually. This alarming statistic underscores the urgent need to comprehend and address the contributing factors to CVDs, which encompass hypertension, diabetes, obesity, and unhealthy lifestyles. In an effort to gain deeper insights into the realm of heart diseases, this project adopts a comprehensive approach that combines manual exploratory data analysis with the application of pandas profiling within a Jupyter Notebook environment hosted on Google Colab. The dataset used in this analysis is sourced from the UCI Heart Disease dataset, which is available through the UCI Machine Learning Repository. Initially containing a total of 76 features or attributes gathered from 303 patients, the dataset was refined to focus exclusively on 14 pertinent features, each offering valuable insights into the prediction of heart diseases.

In this dataset provide a comprehensive explanation of the 14 essential features within the dataset: Age, representing the patient's age in years; Sex, indicating the gender of the patient (1 for male, 0 for female); Chest Pain Type (cp), a categorical variable classifying the type of chest pain experienced into categories (0: asymptomatic, 1: atypical angina, 2: non-anginal pain, 3: typical angina); Resting Blood Pressure (trestbps), denoting the patient's resting blood pressure in mmHg; Serum Cholesterol (chol), presenting the serum cholesterol level in mg/dL; Fasting Blood Sugar (fps), a binary variable signifying whether fasting blood sugar exceeds 120 mg/dL (1 for yes, 0 for no); Resting ECG (restecg), a categorical feature describing resting electrocardiogram results, classified as (0: showing probable or definite left ventricular hypertrophy by Estes’ criteria, 1: normal, 2: having ST-T wave abnormality); Maximum Heart Rate Achieved (thalach), indicating the highest heart rate achieved during testing; Exercise Induced Angina (exang), a binary attribute detecting the presence of exercise-induced angina (1 for yes, 0 for no); ST Depression Induced by Exercise Relative to Rest (oldpeak), quantifying ST depression caused by exercise relative to rest; Slope of the Peak Exercise ST Segment (slope), categorized as (0: downsloping, 1: flat, 2: upsloping); Number of Major Vessels (ca), ranging from 0 to 3, though some entries incorrectly listed as 4 were rectified to NaN; Thalassemia (thal), categorized as (1 = normal, 2 = fixed defect, 3 = reversible defect), with two instances of '0' corrected to NaN; and finally, the Target feature, a binary outcome variable signifying the presence (1) or absence (0) of heart disease.

During the data preprocessing phase, additional steps were undertaken to enhance the quality of the dataset. Missing values were visualized using the Missingno library, with these gaps in the data represented by horizontal lines. To handle these missing data points effectively, NaN values were replaced with the medians of their respective columns. Subsequently, a thorough statistical analysis was conducted. This analysis encompassed the assessment of the range (min-max) for categorical variables such as Gender (Sex), Chest Pain Type (CP), Fasting Blood Sugar (FBS), among others. Additionally, continuous variables were examined, focusing on their mean, standard deviation, and quartile values. Remarkably, the distribution of the target variable revealed a higher prevalence of heart disease cases when compared to individuals without the condition. Further exploration was carried out, revealing that the age distribution exhibited a normal pattern, with a majority of patients falling within the age range of their 50s to 60s. On average, patients were around 54 years old, with a standard deviation of approximately ±9.08. An analysis of gender distribution concerning the target variable uncovered a higher incidence of heart disease among males. However, examining the distribution of chest pain types (CP) raised intriguing questions, particularly regarding the relatively higher number of healthy subjects experiencing typical angina.

The distribution of fasting blood sugar (FBS) indicated that while the true class (diabetic) had a lower count, a significant number of heart disease patients did not have diabetes. This observation suggested that FBS might not serve as a strong differentiating feature for predicting heart disease in this dataset. Moreover, the distribution of the slope variable revealed normal distributions for age, resting blood pressure (trestbps), and almost for serum cholesterol (chol). Conversely, the variables oldpeak and thalach displayed left-skewed and right-skewed distributions, respectively. To gain a comprehensive view of feature distributions, a pairplot was constructed using Seaborn (sns). This visualization highlighted a linear separation relationship between the oldpeak variable and the presence of heart disease. Additionally, a mild separation was observed between thalach and heart disease. Delving deeper into the data, a correlation matrix analysis was conducted, revealing valuable insights into the relationships between various factors and the likelihood of heart disease. Notably, Chest Pain Type, Maximum Heart Rate (Max Heart Rate), Slope, Exercise Induced Angina, Oldpeak, and Vessel Color exhibited strong correlations with the presence of heart disease. Conversely, Fasting Blood Sugar displayed the lowest correlation, suggesting that variations in blood sugar levels may not strongly correlate with heart disease in this dataset. One particularly intriguing finding was the robust correlation identified between the slope of the peak exercise ST segment and OldPeak. This correlation hinted at a potential clinical connection between changes in the ST segment slope during exercise and the magnitude of ST segment depression. Further investigation holds the promise of shedding light on the diagnostic significance of this correlation within the context of heart disease detection.

A screenshot of a graph

Description automatically generated

Figure. 2. Correlation Matrix Pima Indian Diabetes Dataset

A chart with numbers and text

Description automatically generated with medium confidence

Figure. 3. Correlation Matrix Heart Disease UCI Dataset

**2.3 Data Cleaning and Pre-Processing**

Data cleaning and preprocessing play a critical role in preparing the dataset for analysis or machine learning task. The dataset contains valuable information related to diagnosis, so it might require several steps to ensure its accuracy and suitability. In this study, the data cleaning and preprocessing applied by handling the missing values by imputing them using the mean value, Removing Duplicates to avoid inflating the importance of certain instances, Standardization and Normalization to ensure that features have similar scales, preventing any one feature from dominating the analysis, and Data Formatting to ensure that all data types are correctly represented, and there are no formatting issues is crucial.

**2.4 Modelling**

**2.4.1 Linear based model**

Models such as Logistic Regression and SVM operate on the basis of class separation through an understanding of the boundary lines or hyperplanes in the feature space. Logistic Regression uses the logistic function to model class probabilities, whereas SVM looks for hyperplanes that maximize the margin between classes.

**Logistic regression**

Logistic Regression is a statistical technique that installs a linear model to explain the relationship between logit characteristics and one or more independent variables in order to forecast the completion of the categorization of binary results[18]. Logistic Regression is a simple yet popular approach in the Machine Learning community [19]. Used to investigate the association between a binary answer variable and one or more predictor variables (which may be continuous or categorical) [20]. Logistic Regression is used to model the relationship between categorical and covariate response variables. In a logistic model, there is a linear combination of the independent variables with the log probability of an event. Given the covariate values, Logistic Regression assesses the possibility of binary variable features [21]. In order to make precise predictions, logistic regression can also be effectively integrated with other algorithms such as genetic algorithms, evolutionary generalized radial basis functions, and non-parametric models[22].

**Support vector machine (SVM)**

An optimum margin-based classification method in machine learning is called a support vector machine (SVM). SVM, known for its distinctive margin optimization, starts as a linear binary classifier. It has since been extended to encompass nonlinear data through the application of Kernels and to handle multiclass scenarios using a variety of algorithms, including one-to-one, one-to-another, Crammer Singer SVM, Weston Watkins SVM, and directed acyclic graph SVM (DAGSVM). These nuances underscore the dichotomy between linear SVMs with linear kernels and nonlinear SVMs featuring nonlinear kernels [23]. Linear SVM prove particularly adept in tackling high-dimensional datasets across diverse applications, such as document classification, word disambiguation, and drug design. Their efficacy is highlighted by their test accuracy, which rivals that of non-linear SVM, coupled with a distinct advantage in terms of faster and more efficient training. As SVM have evolved since their inception, scholars have contributed a spectrum of issue formulations, remedies, and innovative development methodologies. Furthermore, the rapid evolution of technology has introduced a new challenge for SVM – the handling of 'Big Data' or large-scale datasets [24].

Noteworthy advantages characterize the utilization of the SVM model [25] such as

• Leveraging only pertinent data points as supporting vectors to delineate hyperplanes, effectively mitigating outlier effects and enhancing system efficiency.

• Demonstrating heightened efficacy in small non-linear datasets and high-dimensional spaces, outperforming neural networks.

• Exhibiting resilience against overfitting, a common pitfall in machine learning models.

• Excelling when dimensional characteristics outweigh training samples, presenting a formidable solution for cases of extreme binary classification.

• Excelling in separable datasets, a realm where neural networks often struggle.

However, these advantages are tempered by challenges. The use of SVMs introduces additional computation time when handling substantial datasets. Moreover, their efficacy relies heavily on judicious kernel and hyperparameter selection. SVMs may falter when confronted with data characterized by noise and overlapping class boundaries, leading to suboptimal performance [25].

**2.4.2 Tree based model.**

Models such as Random Forest and Decision Tree build a tree-like structure to divide the feature space and classify data based on the decisions taken at each node. The Decision Tree is a basic model that divides by features, while the Random Forest uses an ensemble of trees to improve accuracy and overcome overfitting.

**Random forest**

The Random Forest (RF) algorithm is a supervised training algorithm that constructs a "forest" consisting of multiple decision trees trained using the "bagging" method. The fundamental idea behind the bagging method is that the combination of multiple learning models enhances the overall capacity [26]. Decision trees, which are an accumulation of numerous weaker learners, give it its predictive capacity [27]. The bagging method is used to train RF, which is regarded as a reliable classification methodology. The creation of a bootstrapped dataset is the initial stage in the training process. In the training process, a specified number of subsets of variables are used. Decision trees are formed, and a class is assigned to a set of data by selecting the decision tree with the most votes [28].

In the random forest building method, a decision tree is initially built for each training subset, and then the process is repeated many times to produce a "forest" of decision trees, where each decision tree develops randomly without being trimmed throughout the development phase [29]. By interrupting the greedy splitting algorithm during tree construction, Random Forest (RF) surpasses bagged decision trees. It achieves this by randomly selecting a subset of input attributes as split points. Unlike traditional classification trees, RF evaluates the bifurcation breakdown only on a randomly chosen subset, similar to the bagging approach. This method employs "simultaneous ensembling," where multiple decision tree models run in parallel on different sub-samples of the dataset. The final conclusion or result is determined through voting. The size of the random variable is predetermined. As a result, the RF algorithm minimizes the problem of overfitting and improves prediction accuracy with controlled variance [30].

**Decision tree**

The Decision Tree (DT) algorithm, which is a member of the supervised learning class of algorithms, is frequently used to solve classification issues, but it may also be applied to situations of classifying and regressing. It comprises of inner nodes that reflect the branch structures, dataset, indicating the algorithm's decision, and each leaf node that represents a result [31]. In order to evaluate the generalizability of the model and prevent over-fitting, DT analysis performed ten-fold cross validation. Accuracy, area under the receiver operating characteristic (AUROC), and confidence intervals may all be determined using this method [32].

The entropy value, which always ranges between 0 and 1, is used to assess the impurity or unpredictability of a dataset. The value is better if it is equal to 0 while worse if it is equal to 0, i.e., the closer the value is to 0 the better [33]. At each stage, the Decision Tree selects each node by evaluating the highest information gain among all attributes, the evaluation uses the Confusion Matrix technique [34].

A classifier model called a decision tree makes judgments using a tree-like structure. Decision trees are frequently employed in various scientific domains because they perform better than alternative methods [35]. Using the testing data, the generated Decision Tree model is put to the test. Data used for testing is information that has been gathered and produced. The training and testing set of data cannot be identical. Performance of the Decision Tree model is gauged by how accurately it generates output from test data. Comparing the output of the testing data with the real output is how accuracy is determined [36].

**Extreme Gradient Boosting (XGBoost)**

Xtreme Gradient Boosting (XGBoost) is a boosting method, namely a collection of decision trees where the next tree development will depend on the previous tree [37]. The first tree in XGboost is weak in classifying by initializing the probability determined by the researcher and then updating the weights of each tree built to produce a collection of strong classification trees [38].

XGBoost introduces the term regularization in the objective function to prevent overfitting which is defined as follows:

|  |  |
| --- | --- |
|  | (1) |

Where 𝐿(, 𝐹()) is the loss function, 𝑅() denotes the regularization term at iteration time 𝑘 and 𝐶 is a constant term that can be omitted selectively. The regularization term 𝑅() is expressed as:

|  |  |
| --- | --- |
|  | (2) |

Where represents the complexity of the leaves, indicates the number of leaves, indicates the penalty parameter and is the output result of each leaf node. In particular a leaf denotes a predicted category following the classification rules and a leaf node denotes an indivisible tree node. In addition, instead of using first-order derivatives in GBDT, a second-order set of taylor objectives is adopted in XGBoost [39]. If the mean squared error (MSE) is used as the loss function, then the objective function can be derived as follows:

|  |  |
| --- | --- |
|  | (3) |

where 𝑞() denotes a function that assigns data points to the corresponding leaf, and denotes the first and second derivatives of the loss function respectively. The final loss value is calculated based on the sum of all loss values. Since the sample corresponds to the nodes in the decision tree, the final loss value can be determined by adding up the loss values of the leaf nodes. Therefore, the objective function is also expressed as follows:

|  |  |
| --- | --- |
|  | (4) |

where = ∑𝑖∈, = ∑𝑖∈, and 𝐼 denotes all samples in the leaf node 𝑗. In other words, the optimization of the objective function is changed as a case of selecting the minimum of the quadratic function. In addition, due to the introduction of the term regularization, XGBoost has a better ability to fight overfitting problems [40].

**2.4.3 Deep learning model**

**Artificial neural network (ANN)**

In general, the ANN model is a method for data analysis that is primarily inspired by the neural systems of humans and other animals. The statistical data analysis model between the input and output variables is non-linear. Numerous scientific and technological domains, including pattern recognition, process control, and time series forecasting, have employed this concept [41]. A sort of parallel computer system known as an Artificial Neural Network (ANN) is created to analyze and process data in a manner like that of the human brain. Using samples from prior data, it is possible to model, categorize, identify, and predict issues [42]. An artificial neural network, or ANN, is a model that takes its cues from how the human brain is biologically stimulated to process information [43].

ANN can classify in various ways depending on the features of its children. It learns from an example because they have adaptive characteristics. In other words, they can learn from previous data and recognize data patterns that are always changing. Besides, ANN is a non-programmed system, meaning that all outputs or conclusions drawn by the network are based on its experience during the learning/training process. In general, the classification using ANN can be based on [44]:

* Functions designed to perform clustering.
* The degree of connection (partial or full) among the network's neurons.
* A type of learning algorithm is a systematic collection of equations that may update artificial neural networks (ANNs) using both the output from the network and any performance metric.
* Learning rule method that can improve ANN performance.

The artificial neurons are similar to biological neuron cells. Neurons will receive information (input) with a specific weight. A function that processes this input will combine the current weight values together. The activation function of each neuron will then be used to compare the sum of the results with a predetermined threshold value. Neurons process each pattern of input and output data that is provided to ANN. Neuron layers are layers where the neurons are gathered [45].

There are three layers that make up the ANN, and they are as follows:

* Input Layer - Input units refer to the units in the input layer. These input devices get input data patterns that indicate issues from the outside.
* Hidden Layer - Hidden units are the units that make up the hidden layer. where the result cannot be seen immediately.
* Output Layer - The units in the output layer are called output units. The output of this layer is an ANN solution to a problem.

**2.4.4 Probabilistic model**

**Gaussian naive bayes**

Gaussian Naive Bayes is a model of supervised learning algorithms that employ the Bayes theorem under the "naive" assumption that each pair of characteristics is independent. A Naive Bayes classifier count up the possibility that given instance (model) belongs to a certain class. Using the naive independence assumption [46],

Given a collection of input attribute value values, this classifier calculates the probability of occurrence of the output feature. The likelihood of each value in relation to the feature will be determined internally. We take into consideration the value of the output attribute that has the highest probability after computing the probabilities of each value. For big datasets, the naive Bayes model is incredibly straightforward, quick, and effective for binary classification [47].

Bayes' theorem enables us to formulate the conditional probability P(y|X) as a combination of simpler probabilities given an instance X and a class target y:

|  |  |
| --- | --- |
|  | (5) |

Where, P(X) is constant for a particular instance, the formula used to classify the sample [48]:

|  |  |
| --- | --- |
|  | (6) |

The Gaussian Naive Bayes is a Naive Bayes variation that uses continuous data for classification and implements the Gaussian normal distribution. GNB leverages the Gaussian distribution to effectively model the probability distributions of continuous attributes. Notably, this choice is particularly apt when the dataset under examination encompasses a range of continuous variables. The GNB algorithm is implemented in the following manner, as delineated by prior research [49]:

* Probability Density Estimation: GNB undertakes the estimation of probability density functions (PDFs) for each class within the dataset. This entails the computation of mean and standard deviation parameters for each continuous feature, enabling the formulation of Gaussian distributions for these features within each class.
* Prior Probabilities: The algorithm further computes the prior probabilities associated with each class, signifying the prior likelihood of each class within the dataset. Typically, this involves a frequency-based calculation derived from the training dataset.
* Posterior Probabilities: When confronted with new input samples characterized by continuous features, GNB proceeds to evaluate the likelihood of the feature values under the Gaussian distribution associated with each class. Bayes' theorem is then employed to derive the posterior probabilities for each class, conditioned on the observed features.
* Prediction: To effectuate predictions, GNB selects the class exhibiting the highest posterior probability as the forecasted class for the input sample. In essence, it ascribes the class label that maximizes the conditional probability of the observed features given the class.

In practice, GNB offers a pragmatic approach for classification tasks involving continuous data, notwithstanding its 'Naive' characterization due to the simplifying assumption of feature independence within each class. The efficacy of GNB is often observed, particularly when the Gaussian distribution assumption aligns with the underlying data distribution. This algorithm holds notable utility for modestly-sized datasets and can be applied to multiclass classification scenarios.

**2.4.5 Evaluation model**

In the evaluation of classification models, the scrutiny of model performance emerges as an imperative endeavor, fundamental to ensuring the credibility and robustness of predictive models. In this study, we subject seven prominent predictive models (SVM, Logistic Regression, XGBoost, Gaussian Naive Bayes, Random Forest, Artificial Neural Networks, and Decision Tree) to meticulous evaluation and analysis. Our objective is to discern the optimal model by scrutinizing and comparing key performance metrics, including accuracy, precision, and recall.

Accuracy, a central measure, assumes a pivotal role. Accuracy offers a comprehensive assessment of the correctness of a classification model's predictions. It elegantly quantifies the ratio of correctly predicted instances to the total number of instances within the dataset. Equation (7) encapsulates the precise formula for accuracy calculation. It is crucial to underscore that while a high accuracy value signifies the model's proficiency in rendering a substantial proportion of correct predictions, prudent consideration is warranted. Sole reliance on accuracy may prove misleading, particularly in scenarios characterized by dataset imbalance.

|  |  |
| --- | --- |
|  | (7) |

Precision represents a pivotal metric, specifically attuned to assessing the accuracy of positive predictions generated by the model. It meticulously quantifies the ratio of true positive predictions, which denote accurate positive identifications, among the entirety of positive predictions. The precise formula for precision computation is elegantly encapsulated in Equation (8). Elevated precision values serve as a compelling indicator, signifying that when the model anticipates a positive outcome, it possesses a notably higher likelihood of being correct. This metric assumes pronounced relevance, particularly in scenarios where the costs associated with false positives hold considerable significance. Furthermore, precision holds a paramount role when demanding unwavering confidence in the precision of positive predictions becomes imperative.

|  |  |
| --- | --- |
|  | (8) |

Recall, also recognized as sensitivity or the true positive rate, serves as a critical gauge of the model's effectiveness in apprehending genuine positive instances within the dataset. It precisely quantifies the ratio of true positives, those instances correctly identified as positive, among the entire set of actual positive instances. The mathematical expression to compute recall is formally denoted as Equation (9). Elevated recall values signify a model's proficiency in capturing a substantial fraction of the authentic positive instances. This metric assumes profound significance in contexts where the ramifications of overlooking positive cases are profound. Notably, in domains where missing positive instances could bear substantial costs or where a thorough detection of positive cases is imperative, a high recall becomes an indispensable attribute for the model.

|  |  |
| --- | --- |
|  | (9) |

In the context of this research, these metrics collectively serve as essential tools to evaluate the performance of various classification models across different datasets. By carefully considering accuracy, precision, and recall, we gain a holistic understanding of how well each model aligns with the specific objectives and challenges of the classification task at hand. These metrics enable us to make informed decisions regarding model selection and highlight areas where model performance can be further enhanced. Through rigorous analysis and interpretation of these evaluation metrics, our research contributes not only to advancing the field of classification modeling but also to providing valuable insights and recommendations for practical applications in domains where accurate categorization is of paramount importance.

1. **RESULT ANALYSIS, COMPARISON, FINDINGS, AND RECOMMENDATION**

After the comprehensive modeling of the three distinct datasets utilizing the seven classification models, as summarized in Table 1. Strikingly, the analysis unveils Random Forest as the standout performer among the assessed models for this dataset. Exhibiting an impressive accuracy of 81.81%, Random Forest simultaneously achieves precision and recall values of 69.38% and 72.34%, respectively. Following Random Forest, XGBoost emerges as the next notable contender, distinguishing itself with a higher recall value than Random Forest. The accuracy, precision, and recall results for XGBoost stand at 81.16%, 67.30%, and 74.46%, respectively. A striking observation arises from the Support Vector Machine (SVM), which notably surpasses other models with exceptional precision and recall values, both at 75%. This achievement underscores the SVM's proficiency in accurately identifying positive cases while effectively minimizing false positives and false negatives. Such a balanced performance is particularly valuable in practical applications like medical diagnosis, anomaly detection, and information retrieval, where precise identification of relevant cases is paramount while maintaining a low rate of false alarms remains crucial. However, it should be noted that the SVM showed a lower overall accuracy of 76.19% compared to other models, indicating that its predictions may not be uniformly correct for both positive and negative cases.

Table 1. Pima Indian Diabetes Dataset Classification

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| Random Forest | 81.81% | 69.38% | 72.34% |
| XGBoost | 81.16% | 67.30% | 74.46% |
| Gaussian Naive Bayes | 78.57% | 65.90% | 61.70% |
| ANN | 77.27% | 75% | 75% |
| SVM | 76.19% | 71.67% | 53.09% |
| Logistic Regression | 74.46% | 68.33% | 50.62% |
| Decision Tree | 66.23% | 50% | 51.92% |

Table 2 presents the results obtained from our classification analysis performed on the UCI Heart Disease Dataset. Notably, our findings reveal the exceptional performance of three prominent classification models Random Forest, XGBoost, and Decision Tree. It is particularly intriguing that these three models yield identical and remarkable accuracy, precision, and recall values, all attaining 99.02%, 100.00%, and 98.34%, respectively. Within the scope of our study, the collective achievement of such elevated accuracy, precision, and recall metrics underscores the exceptional prowess of these models in accurately classifying both positive and negative cases, while simultaneously minimizing the incidence of both false positives and false negatives. This outcome holds significant practical implications, signifying the deployment of well-balanced models capable of reliably identifying a substantial portion of actual positive cases (high recall), providing positive predictions with unwavering confidence (high precision), and ensuring an overall high level of prediction correctness (high accuracy), thereby maintaining a substantial proportion of both true positives and true negatives.

Table 2. Heart Disease UCI Dataset Classification

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** |
| Random Forest | 99.02% | 100.00% | 98.34% |
| XGBoost | 99.02% | 100.00% | 74.46% |
| Gaussian Naive Bayes | 82.46% | 85.08% | 85.08% |
| ANN | 83.11% | 81.46% | 92% |
| SVM | 68.83% | 74.85% | 70.71% |
| Logistic Regression | 82.46% | 85.47% | 84.53% |
| Decision Tree | 99.02% | 100.00% | 98.34% |

Findings:

* Performance of Random Forest: Random Forest consistently produces the best results for both datasets. This indicates that the Random Forest model is a strong choice for classification problems in datasets with unbalanced and multi-feature characteristics.
* Ensemble Learning Capability: Random Forest is a model that utilizes ensemble learning techniques, where it combines multiple decision trees to make a final decision. This allows Random Forest to overcome overfitting because each tree only looks at a small portion of the data, and their results are collectively integrated. This also helps reduce variance in prediction outcomes.
* Reliability Against Outliers: Random Forest is also generally more robust against outliers compared to some other models. This is particularly beneficial for datasets that may contain unusual or inconsistent data points.
* Feature Management Capability: Random Forest automatically performs feature selection, which can reduce noise in the dataset and help improve the model's accuracy.

Recommendations:

* Our analysis yields a strong recommendation for the utilization of tree-based models, notably Random Forest, when implementing classification tasks on datasets characterized by multiple features and class imbalance.
* Feature Engineering: For future research, further exploration of the dataset's features can be conducted to identify the most influential features and eliminate irrelevant ones, which is expected to significantly enhance the model's performance.
* Hyperparameter Optimization: For Random Forest and other models, it's essential to perform hyperparameter optimization to ensure optimal performance. Techniques like cross-validation can be used to determine the best parameters for each model.

Handling Unbalanced Data: If the dataset suffers from class imbalance issues, techniques like oversampling or undersampling can be applied to improve the model's performance in predicting minority classes.

**4. CONCLUSION**

In summary, the research findings underscore the consistent and outstanding performance of the Random Forest model across both the Pima Indian Diabetes and UCI Heart Disease datasets. With an average accuracy of approximately 90.42% for the Pima Indian Diabetes dataset and an impressive 99.02% for the UCI Heart Disease dataset, Random Forest, XGBoost, and Decision Tree exhibit equivalent excellence in accurately classifying data across diverse scenarios. Furthermore, the model's average precision of approximately 84.69% and 100.00% for the respective datasets highlights its exceptional ability to minimize false positives and consistently deliver precise results.

Additionally, the model's recall values of approximately 72.34% for the Pima Indian Diabetes dataset and 98.34% for the UCI Heart Disease dataset highlight its exceptional ability to detect positive cases within these datasets, especially in the context of medical diagnosis. These findings underscore Random Forest as a robust choice for datasets characterized by imbalanced classes and multiple features.

Furthermore, Random Forest's utilization of ensemble learning techniques contributes to its reliability by mitigating overfitting and reducing prediction outcome variance. The model's resilience against outliers further enhances its suitability for datasets containing unusual or inconsistent data points. Its automatic feature selection capability not only improves accuracy but also simplifies the model-building process.

In light of these findings, the research recommends a few key actions. First, future research should delve into feature engineering to identify influential features and eliminate irrelevant ones, thereby enhancing the model's overall performance. Second, it is essential to conduct hyperparameter optimization for Random Forest and other models using techniques such as cross-validation to ensure optimal model configurations. Finally, for datasets with class imbalance issues, implementing oversampling or undersampling techniques can significantly improve the model's ability to predict minority classes. In conclusion, Random Forest emerges as a robust and superior choice for handling the specific characteristics of datasets with abundant features and imbalanced classes, and the recommendations provided offer avenues for further model enhancement and research.

**ReferenSI**

[1] S. Vishwakarma and B. Chilwal, “Detection and Classification of Leaf Blast Disease using Decision Tree Algorithm in Rice Crop,” in *Mathematics and Computer Science Volume 2*, Wiley, 2023, pp. 49–58. doi: 10.1002/9781119896715.ch4.

[2] M. Ozcan and S. Peker, “A classification and regression tree algorithm for heart disease modeling and prediction,” *Healthcare Analytics*, vol. 3, p. 100130, Nov. 2023, doi: 10.1016/j.health.2022.100130.

[3] V. Baviskar, M. Verma, P. Chatterjee, and G. Singal, “Efficient Heart Disease Prediction Using Hybrid Deep Learning Classification Models,” *IRBM*, vol. 44, no. 5, p. 100786, Oct. 2023, doi: 10.1016/j.irbm.2023.100786.

[4] S. M. Ahmad and N. M. Ahmed, “Classification based on event in survival machine learning analysis of cardiovascular disease cohort,” *BMC Cardiovasc Disord*, vol. 23, no. 1, p. 310, Jun. 2023, doi: 10.1186/s12872-023-03328-2.

[5] D. Gupta, K. Attal, and D. Demner-Fushman, “A dataset for medical instructional video classification and question answering,” *Sci Data*, vol. 10, no. 1, p. 158, Mar. 2023, doi: 10.1038/s41597-023-02036-y.

[6] S. Benbelkacem and B. Atmani, “Random Forests for Diabetes Diagnosis,” in *2019 International Conference on Computer and Information Sciences (ICCIS)*, IEEE, Apr. 2019, pp. 1–4. doi: 10.1109/ICCISci.2019.8716405.

[7] M. Pal and S. Parija, “Prediction of Heart Diseases using Random Forest,” *J Phys Conf Ser*, vol. 1817, no. 1, p. 012009, Mar. 2021, doi: 10.1088/1742-6596/1817/1/012009.

[8] V. Chang, J. Bailey, Q. A. Xu, and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” *Neural Comput Appl*, vol. 35, no. 22, pp. 16157–16173, Aug. 2023, doi: 10.1007/s00521-022-07049-z.

[9] Y. Wang, G. Kang, L. Liu, and Q. Huang, “Automatic Assistance Method for Disease Diagnosis Based on a Deep Learning Fusion Model and Chinese Electronic Medical Record,” in *2021 11th International Conference on Bioscience, Biochemistry and Bioinformatics*, New York, NY, USA: ACM, Jan. 2021, pp. 1–8. doi: 10.1145/3448340.3448341.

[10] A. R. B. Alamsyah, S. R. Anisa, N. S. Belinda, and A. Setiawan, “SMOTE and Nearmiss Methods for Disease Classification with Unbalanced Data,” *Proceedings of The International Conference on Data Science and Official Statistics*, vol. 2021, no. 1, pp. 305–314, Jan. 2022, doi: 10.34123/icdsos.v2021i1.240.

[11] H. Symum and J. L. Zayas-Castro, “Prediction of Chronic Disease-Related Inpatient Prolonged Length of Stay Using Machine Learning Algorithms,” *Healthc Inform Res*, vol. 26, no. 1, p. 20, 2020, doi: 10.4258/hir.2020.26.1.20.

[12] H. Hairani and D. Priyanto, “A New Approach of Hybrid Sampling SMOTE and ENN to the Accuracy of Machine Learning Methods on Unbalanced Diabetes Disease Data,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 8, 2023, doi: 10.14569/IJACSA.2023.0140864.

[13] J. Abdollahi and B. Nouri-Moghaddam, “A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation,” *Iran Journal of Computer Science*, vol. 5, no. 3, pp. 229–246, Sep. 2022, doi: 10.1007/s42044-022-00104-x.

[14] T. Milo and A. Somech, “Automating Exploratory Data Analysis via Machine Learning: An Overview,” in *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, New York, NY, USA: ACM, Jun. 2020, pp. 2617–2622. doi: 10.1145/3318464.3383126.

[15] A. Sharma, M. K. Sharma, and R. Kr. Dwivedi, “Exploratory data analysis and deception detection in news articles on social media using machine learning classifiers,” *Ain Shams Engineering Journal*, vol. 14, no. 10, p. 102166, Oct. 2023, doi: 10.1016/j.asej.2023.102166.

[16] J. R. Gautney, “A new approach to exploratory data analysis in hominin phylogenetic reconstruction,” *J Hum Evol*, vol. 182, p. 103412, Sep. 2023, doi: 10.1016/j.jhevol.2023.103412.

[17] V. Bhavani, A. K. Sirivarshitha, K. Sravani, and K. S. Priya, “Music Classification and Mental Health Analysis using Exploratory Data Analysis,” in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, Mar. 2023, pp. 555–561. doi: 10.1109/ICIDCA56705.2023.10099605.

[18] S. Dreiseitl and L. Ohno-Machado, “Logistic regression and artificial neural network classification models: a methodology review,” *J Biomed Inform*, vol. 35, no. 5–6, pp. 352–359, Oct. 2002, doi: 10.1016/S1532-0464(03)00034-0.

[19] M. Abedini, A. Bijari, and T. Banirostam, “Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree, Logistic Regression and Neural Network,” *IJARCCE*, vol. 9, no. 7, pp. 1–4, Jul. 2020, doi: 10.17148/IJARCCE.2020.9701.

[20] R. Y. Yagoub, H. Y. Eledum, and A. A. Yassin, “Factors Affecting the Academic Tripping at University of Tabuk Using Logistic Regression,” *Sage Open*, vol. 13, no. 1, p. 215824402211451, Jan. 2023, doi: 10.1177/21582440221145118.

[21] R. D. Joshi and C. K. Dhakal, “Predicting Type 2 Diabetes Using Logistic Regression and Machine Learning Approaches,” *Int J Environ Res Public Health*, vol. 18, no. 14, p. 7346, Jul. 2021, doi: 10.3390/ijerph18147346.

[22] M. Yaseliani and M. Khedmati, “Prediction of Heart Diseases Using Logistic Regression and Likelihood Ratios,” *International Journal of Industrial Engineering & Production Research*, vol. 34, no. 1, pp. 1–15, 2023, doi: 10.22068/IJIEPR.34.1.5.

[23] E. Z. Astuti, C. A. Sari, M. Syabilla, H. Sutrisno, E. H. Rachmawanto, and M. Doheir, “Capital Optical Character Recognition Using Neural Network Based on Gaussian Filter,” *Scientific Journal of Informatics*, vol. 10, no. 3, pp. 261–270, Jul. 2023, doi: 10.15294/sji.v10i3.43438.

[24] V. K. Chauhan, K. Dahiya, and A. Sharma, “Problem formulations and solvers in linear SVM: a review,” *Artif Intell Rev*, vol. 52, no. 2, pp. 803–855, Aug. 2019, doi: 10.1007/s10462-018-9614-6.

[25] D. A. Otchere, T. O. Arbi Ganat, R. Gholami, and S. Ridha, “Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models,” *J Pet Sci Eng*, vol. 200, p. 108182, May 2021, doi: 10.1016/j.petrol.2020.108182.

[26] P. Y. Taser, “Application of Bagging and Boosting Approaches Using Decision Tree-Based Algorithms in Diabetes Risk Prediction,” in *The 7th International Management Information Systems Conference*, Basel Switzerland: MDPI, Mar. 2021, p. 6. doi: 10.3390/proceedings2021074006.

[27] J. Hu and S. Szymczak, “A review on longitudinal data analysis with random forest,” *Brief Bioinform*, vol. 24, no. 2, Mar. 2023, doi: 10.1093/bib/bbad002.

[28] M. Daviran, M. Shamekhi, R. Ghezelbash, and A. Maghsoudi, “Landslide susceptibility prediction using artificial neural networks, SVMs and random forest: hyperparameters tuning by genetic optimization algorithm,” *International Journal of Environmental Science and Technology*, vol. 20, no. 1, pp. 259–276, Jan. 2023, doi: 10.1007/s13762-022-04491-3.

[29] J.-M. Zhu, Y.-G. Geng, W.-B. Li, X. Li, and Q.-Z. He, “Fuzzy Decision-Making Analysis of Quantitative Stock Selection in VR Industry Based on Random Forest Model,” *Journal of Function Spaces*, vol. 2022, pp. 1–12, Jun. 2022, doi: 10.1155/2022/7556229.

[30] E. S. Mohamed, T. A. Naqishbandi, S. A. C. Bukhari, I. Rauf, V. Sawrikar, and A. Hussain, “A hybrid mental health prediction model using Support Vector Machine, Multilayer Perceptron, and Random Forest algorithms,” *Healthcare Analytics*, vol. 3, p. 100185, Nov. 2023, doi: 10.1016/j.health.2023.100185.

[31] M. Bansal, A. Goyal, and A. Choudhary, “A comparative analysis of K-Nearest Neighbor, Genetic, Support Vector Machine, Decision Tree, and Long Short Term Memory algorithms in machine learning,” *Decision Analytics Journal*, vol. 3, p. 100071, Jun. 2022, doi: 10.1016/j.dajour.2022.100071.

[32] A. Elhazmi *et al.*, “Machine learning decision tree algorithm role for predicting mortality in critically ill adult COVID-19 patients admitted to the ICU,” *J Infect Public Health*, vol. 15, no. 7, pp. 826–834, Jul. 2022, doi: 10.1016/j.jiph.2022.06.008.

[33] B. Charbuty and A. Abdulazeez, “Classification Based on Decision Tree Algorithm for Machine Learning,” *Journal of Applied Science and Technology Trends*, vol. 2, no. 01, pp. 20–28, Mar. 2021, doi: 10.38094/jastt20165.

[34] D. Sisodia and D. S. Sisodia, “Prediction of Diabetes using Classification Algorithms,” *Procedia Comput Sci*, vol. 132, pp. 1578–1585, 2018, doi: 10.1016/j.procs.2018.05.122.

[35] P. Sornsuwit, P. Jundahuadong, and S. Pongsakornrungsilp, “A New Efficiency Improvement of Ensemble Learning for Heart Failure Classification by Least Error Boosting,” *Emerging Science Journal*, vol. 7, no. 1, pp. 135–146, Nov. 2022, doi: 10.28991/ESJ-2023-07-01-010.

[36] M. Abdurohman, A. G. Putrada, and M. M. Deris, “A Robust Internet of Things-Based Aquarium Control System Using Decision Tree Regression Algorithm,” *IEEE Access*, vol. 10, pp. 56937–56951, 2022, doi: 10.1109/ACCESS.2022.3177225.

[37] T. Chen and T. He, “xgboost: eXtreme Gradient Boosting,” 2023.

[38] R. Mitchell and E. Frank, “Accelerating the XGBoost algorithm using GPU computing,” *PeerJ Comput Sci*, vol. 3, p. e127, Jul. 2017, doi: 10.7717/peerj-cs.127.

[39] T. Chen and C. Guestrin, “XGBoost,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.

[40] H. Jiang, Z. He, G. Ye, and H. Zhang, “Network Intrusion Detection Based on PSO-Xgboost Model,” *IEEE Access*, vol. 8, pp. 58392–58401, 2020, doi: 10.1109/ACCESS.2020.2982418.

[41] Y. Chen *et al.*, “Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential,” *Geocarto Int*, vol. 37, no. 19, pp. 5564–5584, Oct. 2022, doi: 10.1080/10106049.2021.1920635.

[42] M. Vishnupriyan and R. Annadurai, “A study on the macro-properties of PCB fiber-reinforced concrete from recycled electronic waste and validation of results using RSM and ANN,” *Asian Journal of Civil Engineering*, vol. 24, no. 6, pp. 1667–1680, Sep. 2023, doi: 10.1007/s42107-023-00595-4.

[43] S. Agatonovic-Kustrin and R. Beresford, “Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research,” *J Pharm Biomed Anal*, vol. 22, no. 5, pp. 717–727, Jun. 2000, doi: 10.1016/S0731-7085(99)00272-1.

[44] I. A. Basheer and M. Hajmeer, “Artificial neural networks: fundamentals, computing, design, and application,” *J Microbiol Methods*, vol. 43, no. 1, pp. 3–31, Dec. 2000, doi: 10.1016/S0167-7012(00)00201-3.

[45] S. Shanmuganathan, “Artificial Neural Network Modelling: An Introduction,” 2016, pp. 1–14. doi: 10.1007/978-3-319-28495-8\_1.

[46] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, “Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes,” *PLoS One*, vol. 9, no. 1, p. e86703, Jan. 2014, doi: 10.1371/journal.pone.0086703.

[47] V. Sai Krishna Reddy, P. Meghana, N. V Subba Reddy, and B. Ashwath Rao, “Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers,” *J Phys Conf Ser*, vol. 2161, no. 1, p. 012015, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012015.

[48] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, “Sequence Based Prediction of DNA-Binding Proteins Based on Hybrid Feature Selection Using Random Forest and Gaussian Naïve Bayes,” *PLoS One*, vol. 9, no. 1, p. e86703, Jan. 2014, doi: 10.1371/journal.pone.0086703.

[49] V. Sai Krishna Reddy, P. Meghana, N. V Subba Reddy, and B. Ashwath Rao, “Prediction on Cardiovascular disease using Decision tree and Naïve Bayes classifiers,” *J Phys Conf Ser*, vol. 2161, no. 1, p. 012015, Jan. 2022, doi: 10.1088/1742-6596/2161/1/012015.