Optimizing Celiac Disease Detection Through Dataset Balancing and Deep Learning in a Mobile Application

Amalia Amalia*, Maya Silvi Lydia, Sri Melvani Hardi, Andrew Benedictus Jamesie, and Hasanul Fahmi

***Abstract*—Celiac disease is a rare autoimmune condition that targets healthy tissue in the small intestine due to gluten intolerance. Diagnosis typically involves procedures such as intestinal biopsy, small intestine endoscopy, and blood tests. Several previous studies have developed an early detection system for celiac disease using image data. Past research using numerical and categorical data shows indications of data imbalance in the classification labels, which impacts the accuracy level of the ANN model. This research aims to tackle imbalanced data in celiac disease identification by applying and comparing various synthetic oversampling and undersampling techniques, including Adaptive Synthetic Sampling (ADASYN), Synthetic Minority Oversampling Technique (SMOTE), SMOTE combined with Edited Nearest Neighbors (SMOTE-ENN), and SMOTE combined with Tomek Links (SMOTE-Tomek). The objective is to address class imbalance and identify the most effective resampling method to improve celiac disease classification performance. Among the tested methods, the ANN combined with SMOTE-ENN achieved the best results, with a training accuracy of 98.0421% and a training loss of 0.074591, along with a validation accuracy of 98.4326% and a validation loss of 0.103261. The testing accuracy stood at 99.38%, with a loss of 0.0062, and precision, recall, and F1-score of 99%, highlighting excellent predictive balance, crucial in medical diagnosis where both false positives and false negatives carry significant consequences. The model was successfully converted to TensorFlow Lite and deployed in an Android application, enabling real-time, offline prediction of celiac disease. This integration offers a practical, accessible, and scalable approach for early screening using on-device machine learning.**

***Index Terms*— Celiac Disease, Synthetic Oversampling, Synthetic Undersampling, On-device Machine Learning, Dataset Balancing**

Manuscript received March 10, 2025; revised July 11, 2025.

This work was supported in part by the Lembaga Penelitian Universitas Sumatera Utara (LP USU) under a TALENTA USU research grant in 2023.

Amalia Amalia is an associate professor at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia (corresponding author to provide phone: +62 811-645-554; e-mail: amalia@usu.ac.id).

Maya Silvi Lydia is an associate professor at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia (e-mail: maya.silvi@usu.ac.id).

Sri Melvani Hardi is an assistant professor at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia (e-mail: vani.hardi@usu.ac.id).

Andrew Benedictus Jamesie is a bachelor's graduate at the Faculty of Computer Science and Information Technology, Universitas Sumatera Utara, Medan, Indonesia (e-mail: andrew.jamesie@students.usu.ac.id).

Hasanul Fahmi is a senior lecturer at UNITAR International University, Kuala Lumpur, Malaysia (e-mail: fahmi.zuhri@unitar.my).

I. Introduction

AN autoimmune disorder that affects the small intestine in individuals sensitive to gluten is called celiac disease or coeliac disease (CD) [1]. Gluten, a protein present in grains such as wheat, barley, spelt, kamut, and rye [2], triggers an immune response that leads to damage in the histological tissue lining of the small intestine and various health issues. Diagnosis involves biopsy and intestinal endoscopy to assess the condition of the small intestine. Biopsy extracts and analyzes tissue samples, while endoscopy uses a camera-equipped tube for direct visualization and sample collection [3]. However, endoscopy has limitations like low resolution and high costs. The analysis of endoscopy results in the laboratory poses challenges for doctors, and the procedure itself requires anesthesia and intestinal preparation [4]. Celiac disease can also be diagnosed through a blood test, which examines abnormalities in white blood cells and antibodies. In individuals with celiac disease, there is a substantial increase, ranging from 2 to 6 times, in the production of IgA, IgG, and IgM antibodies [5]. This elevated antibody production causes white blood cells to behave abnormally, resulting in the attack of healthy cells in the body. IgA, primarily found in the respiratory, digestive, and ocular systems, defends the body surface against bacterial and viral threats. IgG, the most abundant antibody in the bloodstream and various tissues, fights pathogens in body fluids and provides long-lasting immunity with a memory of previous infections. IgM serves as the body's initial response to new pathogens, offering swift protection during the early stages of an attack.

Early diagnosis of celiac disease is vital for preventing complications, improving quality of life, and reducing healthcare costs associated with treating advanced symptoms or related conditions. Proactive screening in at-risk populations and timely intervention can substantially impact long-term health outcomes [6]. Blood testing is the most affordable method for the early detection of celiac disease. It is non-invasive, quick, and only requires a simple blood draw, making it significantly less invasive than an endoscopy with biopsy, which involves inserting a scope into the digestive tract under sedation. This approach is particularly beneficial for children and individuals who may be apprehensive about undergoing invasive procedures. Additionally, blood tests are widely available, accessible, and practical in most healthcare settings.

However, a challenge arises due to the limited dataset available for celiac disease in several Asian countries, particularly in Indonesia, Korea, and Taiwan, where the prevalence of this disease is low, do not consistently report cases of celiac disease [7], [8], [9]. Typically, existing celiac disease data comprises endoscopy images or intestinal biopsy results [4], [10], [11]. In this study, numerical and categorical data from the Kaggle Celiac disease (coeliac disease) [12] public dataset platform, initially sourced from the Wageningen University & Research (WUR) Biotechnology Department Lab, are employed. This dataset encompasses 2206 rows of data and 15 columns of features. However, the dataset exhibits an imbalanced class distribution, with significant disparities in the number of samples across diagnostic categories, which may affect the performance of classification models if not adequately addressed.

A previous study conducted by [13] employed an Artificial Neural Network (ANN) approach to develop a model for celiac disease identification using a public dataset [12]. The model achieved an accuracy of 97% with a training loss of 0.1105. Validation results showed a validation accuracy of 96.8% and a loss of 0.1475. Furthermore, the model was deployed as a web-based application to support practical implementation. Nevertheless, the study did not sufficiently address the issue of class imbalance in the dataset, which exhibited significant disparities in the number of samples across classes. The model's accuracy could be further improved by balancing the class distribution properly. In addition, transitioning the deployment to a mobile platform could enhance accessibility and broaden user reach, especially in real-world healthcare settings.

Several previous studies have demonstrated that imbalanced data can significantly affect the performance of machine learning models, often resulting in biased predictions that favor the majority class while overlooking the minority class. This imbalance often results in reduced overall accuracy, particularly in applications where the minority class plays a crucial role in decision-making, such as medical diagnoses or detecting rare events. The study by [14] showed how imbalanced data affects deep learning models, especially in multi-class classification settings. Even some Regularization techniques such as Dropout, Reduction on Plateau, and Early Stopping do not provide significant solutions to class imbalances. Another study by [15] showed improved performance after implementing the dataset-balancing process in detecting skin cancer. However, using only traditional methods, such as image data augmentation by rotation, will increase the dataset size, which will positively impact performance and help balance the dataset used.

This research aims to enhance the quality of the identification model by addressing data imbalance through the use of synthetic oversampling and undersampling techniques, thereby increasing model training and testing accuracy. Additionally, we employ a more advanced approach using deep learning techniques. Deep learning is chosen over traditional machine learning models due to its superior ability to capture complex, non-linear patterns in data, its scalability for future model enhancement, and its compatibility with real-time mobile deployment frameworks such as TensorFlow Lite. To further strengthen the model's reliability, this study also aims to identify the most effective data balancing method by comparing several techniques, including ADASYN, SMOTE, SMOTE-ENN, and SMOTE-Tomek. Furthermore, the final model is implemented in an Android-based mobile application to provide a more personalized, accessible, and user-friendly experience for individuals seeking early screening of celiac disease. Based on the background, our main contributions are as follows:

(a) Revising the dataset for celiac identification through blood tests to ensure a more balanced class distribution.

(b) Enhancing the overall accuracy and robustness of the celiac disease classification model by identifying the optimal resampling method and applying an appropriate deep learning architecture.

(c) Developing a more comprehensive, user-friendly system based on an Android platform.

The structure of this paper is outlined as follows: Section 2 introduces related works, Section 3 details the deep learning method, Section 4 describes synthetic oversampling and undersampling techniques, Section 5 outlines the methodology, Section 6 presents the results and discussion, and Section 7 provides the conclusion and future work.

II. Related Works

Several previous related works have focused on different aspects of celiac disease detection. For instance, a study by [16] proposed a multiple-instance learning-based approach for detecting celiac disease. Unlike our research, this study utilized a dataset comprising scanned images of biopsy results. The study focused on identifying only two classes: tissue with detected celiac disease and normal tissue. Meanwhile, our study classifies cases into six classes: non-celiac, potential, latent, silent, atypical, and typical.

Another study by [17] also focuses on celiac detection using machine learning to interpret small intestinal biopsy images, aiding pathologists in streamlining diagnoses while minimizing bias. This research achieved an accuracy of 88.89% for classifying two types of villous abnormalities based on Hematoxylin and Eosin (H&E)- stained biopsy images. Meanwhile, the classification of red-green-blue (RGB) biopsy images reached 82.92% accuracy and 72% accuracy in multi-class biopsy image classification. Similarly, a study conducted by [18] employed the ResNet-18 Convolutional Neural Network (CNN) algorithm to classify Hematoxylin and Eosin (H&E) celiac histological images, distinguishing normal small intestine controls and duodenal inflammation with an impressive accuracy of 99.7%.

Another previous study focused on dataset balancing, highlighting its importance in improving model performance. Various methods have been employed to achieve a balanced dataset, such as oversampling, undersampling, and hybrid approaches. Oversampling is a process to increase the number of samples in the minority class by duplicating existing data or generating synthetic samples using techniques such as SMOTE (Synthetic Minority Oversampling Technique). Meanwhile, undersampling is a process of reducing the number of

samples in the majority class to match the number of samples in the minority class, thereby ensuring equal representation.

A previous study by [19] analyzed the impact of undersampling and oversampling techniques on the classification using an imbalanced dataset. The results showed that both undersampling with random sampling and oversampling with the SMOTE technique yielded higher accuracy than models trained on the original imbalanced dataset, indicating that addressing class imbalance significantly improves classification performance. Specifically, the study achieved an accuracy of 80.45% using the undersampling technique and 79.36% with the oversampling technique, compared to only 70.55% when no balancing method was applied, using the Random Forest algorithm. Research conducted by [20] using SMOTE to randomly overcome the imbalanced data of depression among women, utilizing the random forest algorithm, also succeeded in increasing accuracy by 3.69% and sensitivity/recall by 69.64%. Meanwhile, in a research study by [21], it was concluded that comparing two techniques to tackle and review the imbalance data problem, the oversampling technique had better performance and obtained higher evaluation metrics with several classification algorithms such as Random Forest, Support Vector Machine, Linear Regression, Decision Tree, Naïve Bayes, etc. on the Santander Customer Transaction Prediction dataset.

An alternative methodology is the hybrid approach, which integrates oversampling and undersampling techniques to optimize class distribution in the dataset while minimizing redundancy or data loss. A previous study by [22] increased the accuracy of the validation set by 8.9% using the SMOTE and Edited Nearest Neighbors (SMOTE-ENN) method to optimize hyperparameters on an unbalanced heart failure dataset. Employing SMOTE for oversampling alongside the Edited Nearest Neighbors algorithm as the undersampling method can enhance the model's accuracy and overall performance. Apart from that, the study carried out by [23], also succeeded in achieving an accuracy of 98.925% using the SMOTE method and 98.919% using the SMOTE-Tomek method to make predictions on a cervical cancer dataset that had been balanced with these methods. A research study carried out by [24] also utilized the SMOTE-ENN technique with an Artificial Neural Network (ANN) to achieve better model performance with 95% overall accuracy in classifying Marburg Virus (MARV) inhibitors. The proposed SMOTE-ENN + ANN hybrid model demonstrated superior accuracy and effectiveness in identifying potential lead molecules against MARV.

Research conducted by [25] identified difficulties in making accurate model predictions due to imbalanced data and high dimensionality of the dataset. Using SMOTE and Principal Component Analysis (PCA) can help overcome these two problems in the bankruptcy binary classification prediction dataset. The same oversampling technique is also employed in the research by [26], which proposes KCSMOTE—a combination of the K-means Center model with the SMOTE method—that effectively overcomes unbalanced fraud data compared to other sampling methods. Using the XGBoost and Random Forest algorithms to validate the effectiveness of the proposed method has resulted in an increase of more than 1% compared to other methods. The same thing was also done by research [27] which successfully solved the problem on 25 imbalanced datasets using the K-means clustering method approach with a combination of SMOTE and Random Under-Sampling Technique (RUS), followed by Tomek-Link-based undersampling and Cluster-based undersampling methods to eliminate the majority of samples from overlapping regions, obtaining an average total AUC metric of 90.9% for all test datasets on the AdaBoost-C4.5 model and 88.5% on the AdaBoost-NB model.

III. Deep Learning

Deep learning represents an advanced subset of machine learning (ML) and artificial intelligence (AI) that utilizes multi-layered neural networks to model and process complex patterns in data. This approach aims to emulate the information-processing mechanisms of the human brain by interconnecting numerous layers of neural networks and artificial neurons [28]. Through the deployment of multiple artificial layers and neurons, deep learning can effectively extract valuable features from data, particularly in applications within the health sector for disease detection [29]. Various machine learning algorithms, including Artificial Neural Networks (ANN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN), exhibit the capability to autonomously learn features from raw data, obviating the need for human domain experts. The ability of these algorithms to learn directly from raw data increases efficiency by conserving time, energy, and resources and also enables accurate classification using the extracted features [30].

Artificial Neural Networks (ANN) serve as mathematical and computational models inspired by the structure and functionality of the human brain. Comprising interconnected artificial neurons, this model learns from data [28]. The architecture of an artificial neural network involves distinct layers: the input layer, hidden layers, and the output layer, all interconnected with weighted connections for each neuron [31]. A perceptron or neuron within an artificial neural network is depicted in Fig. 1. The weighted sum calculation is expressed by (1).

$$\hat{y} = g\left(w_0 + \sum_{i=1}^{n} x_i w_i \right) \tag{1}$$

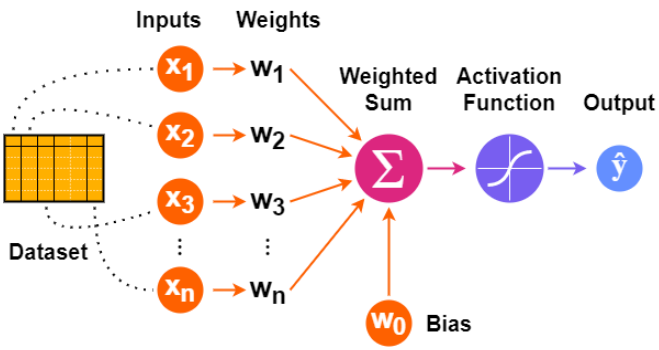


Fig. 1. Model of Neural Network Perceptron.

In this expression, x_i represents the input neuron, w_0 denotes the bias, w_i for the weight, Σ is the weighted sum, g is the activation or non-linearity function, and $\hat{y}$ is the output.

Activation function is a mathematical function that is applied to the output of a neuron in an artificial neural network. The goal is to introduce an element of non-linearity into the model, allowing artificial neural networks to recognize and model complex data patterns. This non-linearity of the function indicates that the output does not come from a simple linear operation on the neuron's input. Some examples of commonly used activation functions include the ReLU (Rectified Linear Unit) activation function, as shown in formula (2), and the Softmax activation function, presented in formula (3).

$$R(z) = \max(0, z) \tag{2}$$

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{n} e^{z_i}} \tag{3}$$

IV. Synthetic Oversampling and Undersampling

One common issue in machine learning, particularly in health datasets, is the presence of imbalanced data classes or labels. This imbalance can lead machine learning models to exhibit bias towards the majority class, hindering their ability to recognize minority classes accurately. Consequently, this imbalance negatively impacts the overall performance and evaluation outcomes of machine learning models [32]. Various techniques exist to address data imbalance, such as undersampling and oversampling, illustrated in Fig. 2 [19]. Undersampling involves reducing the number of samples from the majority class to achieve a more balanced distribution with the minority class. Conversely, oversampling increases the number of samples from the minority class by duplicating existing data or generating synthetic data to balance the class distribution [21].

However, undersampling may lead to a loss of valuable information and the elimination of important data patterns or trends crucial for the machine learning model's learning process [33]. On the other hand, oversampling can result in overfitting, where the model achieves high accuracy on the training data but performs poorly on the test data, thereby increasing the risk of learning noise or irrelevant information [34]. To address these challenges, synthetic oversampling techniques, such as SMOTE, have been developed to enhance the recognition of minority classes and improve overall model performance [35]. SMOTE, introduced by [36], utilizes the K-Nearest Neighbor (KNN) algorithm to generate synthetic data based on the distances to the data's nearest neighbors using Euclidean distance [37]. The classification performance of the imbalanced dataset is enhanced by increasing the data imbalance ratio through the generation of a specific number of synthetic minority samples [38]. The detailed procedure of SMOTE is as follows [39].

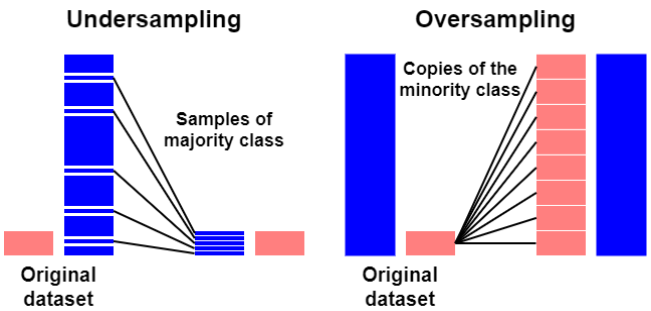


Fig. 2. Illustration of Undersampling and Oversampling Technique.

Step 1: For each minority class sample x_i (i = 1, 2, ..., n), compute its distance from other minority samples using a defined rule to determine its k nearest neighbors within the minority class.

Step 2: Based on the desired oversampling rate, randomly select m nearest neighbors from the set of k nearest neighbors. Denote these selected neighbors as x_{ij} (j = 1, 2, ..., n). Equation (4) is to generate synthetic minority samples p_{ij}, where rand(0,1) is a random value sampled from a uniform distribution within the range [0,1]. This process continues until the dataset reaches the desired imbalance ratio.

$$p_{ij} = x_i + \text{rand}(0,1) \times (x_{ij} - x_i) \tag{4}$$

In addition to SMOTE, the SMOTE-Tomek method was introduced by [40]. It also incorporates a step to remove noisy samples after applying SMOTE, utilizing the Tomek links concept to refine the oversampling process [32].

V. Methodology

This research encompasses various stages, beginning with data understanding, data pre-processing involving oversampling and undersampling techniques, development of a deep learning model, model training, model evaluation to assess its performance, model testing, and finally, on-device machine learning implementation in the Android-based application. The study's methodology is illustrated in Fig. 3.

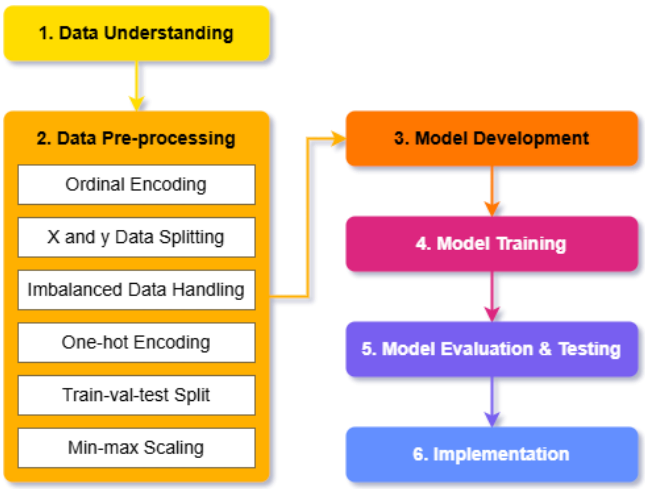


Fig. 3. Methodology.

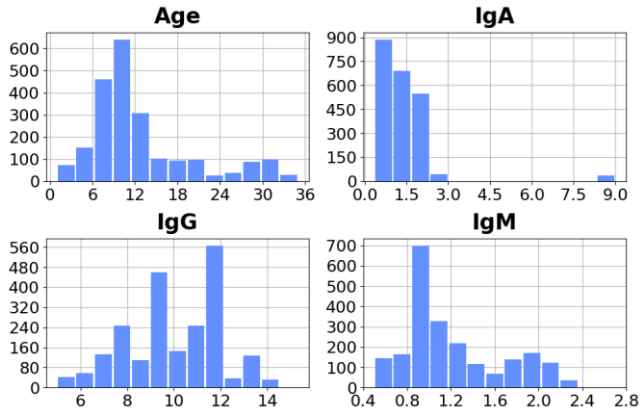


Fig. 4. Univariate Analysis of Numerical Data.

A. *Data Understanding*

The dataset used in this research was sourced from Kaggle's public dataset platform, explicitly focusing on Celiac disease (also known as coeliac disease) [12]. It comprises 2206 rows and 15 columns, encompassing various features such as age, gender, diabetes status, type of diabetes, diarrhea, abdominal symptoms, short stature type, sticky stool, weight loss, immunoglobulin levels (IgA, IgG, IgM), marsh type, celiac disease type, and disease diagnosis.

The findings from the univariate analysis of numerical data reveal a range of values, including age spans from 1 to 35 years old, with the highest frequency occurring around 8 to 10 years old. IgA levels range from 0.34 to 9.0, with the highest being 1.00 to 2.00. While IgM levels were more diverse, ranging from 5.0 to 15.3, with the highest being 12.0, 9.0, and 8.0, and IgM levels were also quite varied, ranging from 0.5 to 2.7, with the highest being 1.00, 1.10, and 2.00. Fig. 4 displays the histogram illustrating the distribution of these numerical variables.

Concurrently, the categorical data univariate analysis in Fig. 5 shows a balanced gender distribution with 1122 men and 1064 women. Most individuals have diabetes, with 1829 having diabetes and the remaining 377 not, with various types, 1663 having Type 1, 125 having Type 2, and the rest having no data. The diarrhea types are evenly distributed, 710 are inflammatory, 773 are fatty, and 723 are watery. Most also have abdominal pain, as many as 1781 people, while 425 others do not. For the short stature variable, 959 individuals were categorized as having Proportionate Short Stature (PSS), 904 as variants, and 343 as having Disproportionate Short Stature (DSS). Regarding the sticky stool variable, 1,820 respondents answered ”yes,” while 386 answered “no.” Similarly, 1,514 individuals reported experiencing weight loss in the weight loss variable, whereas 692 did not.

Different from the marsh type variable, which varies significantly with 7 categories, 350 are none, 313 marsh type 0, 417 marsh type 1, 232 marsh type 2, 232 marsh type 3a, 445 marsh type 3b, and 217 marsh type 3c. The class label has six categories: none, potential, latent, silent, atypical, and typical, with 350, 230, 301, 400, 545, and 380, respectively. It is clearly seen that the cd type (celiac disease type) class label has an unbalanced distribution in each class, so it is necessary to handle this problem using the method that will be explained below.

B. *Data Pre-processing*

Ordinal encoding was applied to convert categorical data into numerical values for variables such as gender, diabetes status, diabetes type, diarrhea type, abdominal symptoms, short stature type, sticky stool, weight loss, and Marsh type.

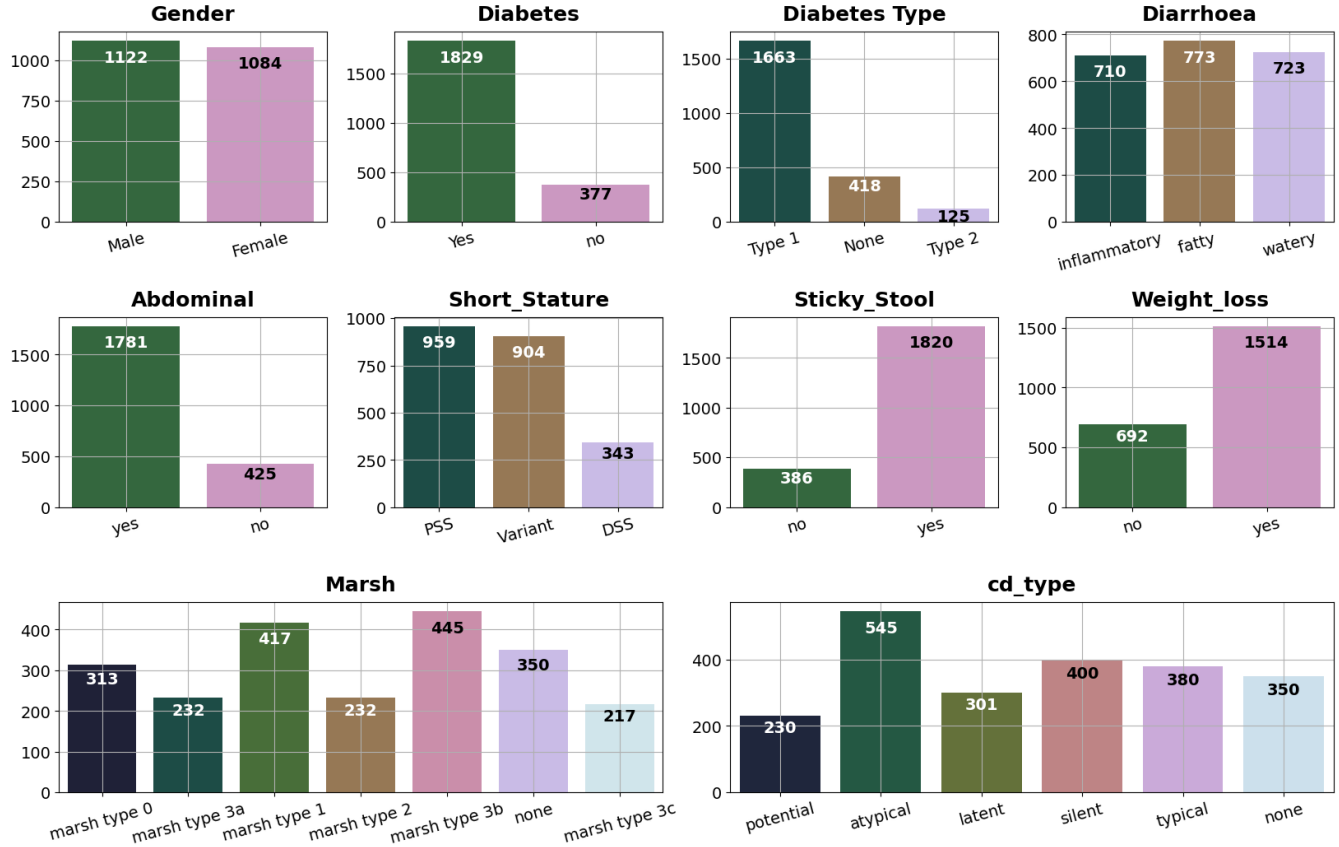


Fig. 5. Univariate Analysis of Categorical Data.

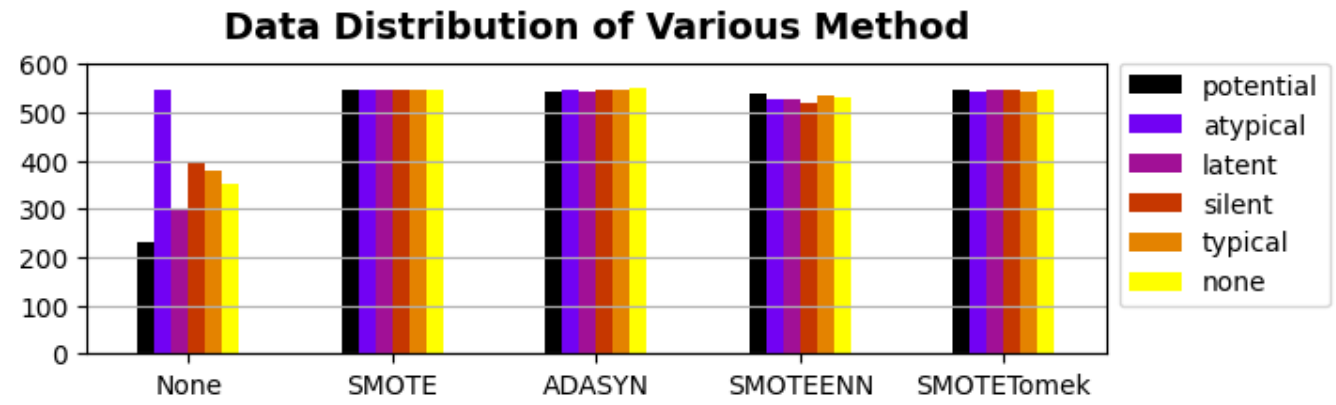


Fig. 6. Comparison of Data Distribution.

The dataset was split into independent variables (X) and a dependent variable (y). Data imbalance was addressed using both oversampling and undersampling techniques, including ADASYN (Adaptive Synthetic Sampling), SMOTE (Synthetic Minority Over-sampling Technique), SMOTE-ENN (SMOTE with Edited Nearest Neighbors), and SMOTE-Tomek (SMOTE with Tomek Links). Table I and Fig. 6 show the data distribution before and after resampling.

TABLE I
COMPARISON OF DATA DISTRIBUTION

Label	No resampling	ADASYN	SMOTE	SMOTE-ENN	SMOTE-Tomek
Typical	380	545	547	536	543
Atypical	545	545	545	528	543
Silent	400	545	546	520	544
Latent	301	545	542	527	544
Potential	230	545	544	540	545
None	350	545	550	535	545
TOTAL	2206	3270	3274	3186	3264

Subsequently, one-hot encoding was applied to the dependent variable, precisely the type of celiac disease, which comprises six classes or data labels: none, typical, atypical, silent, latent, and potential. The dataset was then partitioned into 85% training data, 10% validation data, and 5% testing data, as shown in Table II for each dataset. Following partitioning, normalization scaling was performed using the min-max scaling technique on each data feature to standardize the range of values between 0 and 1, as computed by (5) [41].

TABLE II
COMPARISON OF DATA SPLITTING FOR EACH DATASET

Data Splitting	No resampling	ADASYN	SMOTE	SMOTE-ENN	SMOTE-Tomek
X_train	(1874, 13)	(2782, 13)	(2779, 13)	(2692, 13)	(2771, 13)
X_val	(221, 13)	(328, 13)	(327, 13)	(317, 13)	(327, 13)
X_test	(111, 13)	(164, 13)	(164, 13)	(159, 13)	(164, 13)
y_train	(1874, 6)	(2782, 6)	(2779, 6)	(2692, 6)	(2771, 6)
y_val	(221, 6)	(328, 6)	(327, 6)	(317, 6)	(327, 6)
y_test	(111, 6)	(164, 6)	(164, 6)	(159, 6)	(164, 6)

$$x_{scaled} = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{5}$$

C. Development of Deep Learning Model

The architecture of the deep learning model was constructed using Python with TensorFlow and Keras libraries, employing the Artificial Neural Network (ANN) algorithm. It includes an input layer of 13 neurons, corresponding to the number of input features on the celiac disease dataset. This is followed by three dense layers with 128, 64, and 32 neurons, respectively, utilizing the ReLU activation function to introduce non-linearity and mitigate the vanishing gradient problem. The decreasing number of neurons across layers helps in hierarchical feature extraction while reducing computational complexity. Additionally, to prevent overfitting, dropout layers with a rate of 0.25 were inserted between each dense layer, ensuring a balance between model generalization and performance. The output layer consists of six neurons with a Softmax activation function, which is appropriate for multi-class classification tasks, enabling the model to output probability distributions across six classes. These hyperparameters were chosen by empirical experimentation and best practices to optimize both learning efficiency and model generalization.

D. Model Training

To establish a strong comparative baseline and identify the most effective classification approach for predicting types of celiac disease, various classical machine learning models were initially developed and evaluated. These models included Logistic Regression (with max iteration set to 1000 to ensure convergence), Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Adaptive Boosting (AdaBoost). The selection of these models reflects a diverse set of learning paradigms, ranging from linear (Logistic Regression), probabilistic (Naïve Bayes), and margin-based (SVM) to tree-based (Decision Tree, Random Forest, AdaBoost) and instance-based approaches (KNN), to ensure a comprehensive and representative benchmark.

All models, except for Logistic Regression, were initialized with the default parameters provided by the scikit-learn library to ensure reproducibility and fair

comparison. Each model was trained using pre-processed and resampled datasets obtained through ADASYN, SMOTE, SMOTE-ENN, and SMOTE-Tomek techniques to address class imbalance. The results of these traditional models served as a comparative baseline to evaluate the performance gains offered by the proposed Artificial Neural Network (ANN) architecture.

For the deep learning approach, the model was compiled with the Adam optimizer, chosen for its adaptive learning rate capabilities and efficiency with sparse gradients. The initial learning rate was set at 0.001, a commonly used value that balances convergence speed and stability. The categorical cross-entropy loss function was used for multi-class classification, measuring the difference between predicted probabilities and actual class labels. The accuracy metric was used to evaluate the model's performance. To prevent overfitting and improve generalization, the training process incorporated an early stopping callback function, reducing the risk of excessive training iterations.

Additionally, the model leveraged the ReduceLROnPlateau callback, an adaptive learning rate scheduling technique that dynamically reduces the learning rate when training stagnates. This function monitored the validation loss with a patience of 20 epochs. This step helps the model escape local minima and continue optimizing at a slower pace to fine-tune its performance.

E. Model Evaluation and Testing

The model evaluation aims to assess its ability to comprehend data patterns with traditional machine learning algorithms and the proposed ANN model. During testing, it is evaluated on previously validated data to determine classification accuracy on the data. Evaluation metrics include accuracy and loss function. Additionally, the testing data undergoes analysis using the precision, recall, and F1-score metrics for each data label.

This multi-model evaluation provided a valuable benchmark for identifying models with promising performance characteristics, allowing for a smooth transition to more complex architectures. Based on empirical results and the observed limitations of traditional models in capturing non-linear relationships and complex patterns, an Artificial Neural Network (ANN) model was subsequently developed to enhance predictive performance further.

F. Implementation

In this phase, the model architecture is integrated into an Android-based application developed using the Kotlin programming language. To optimize performance on mobile and embedded devices, the model is converted into a specialized format that supports low-latency inference and low-power consumption. This will enable efficient and seamless execution of prediction tasks within the Android

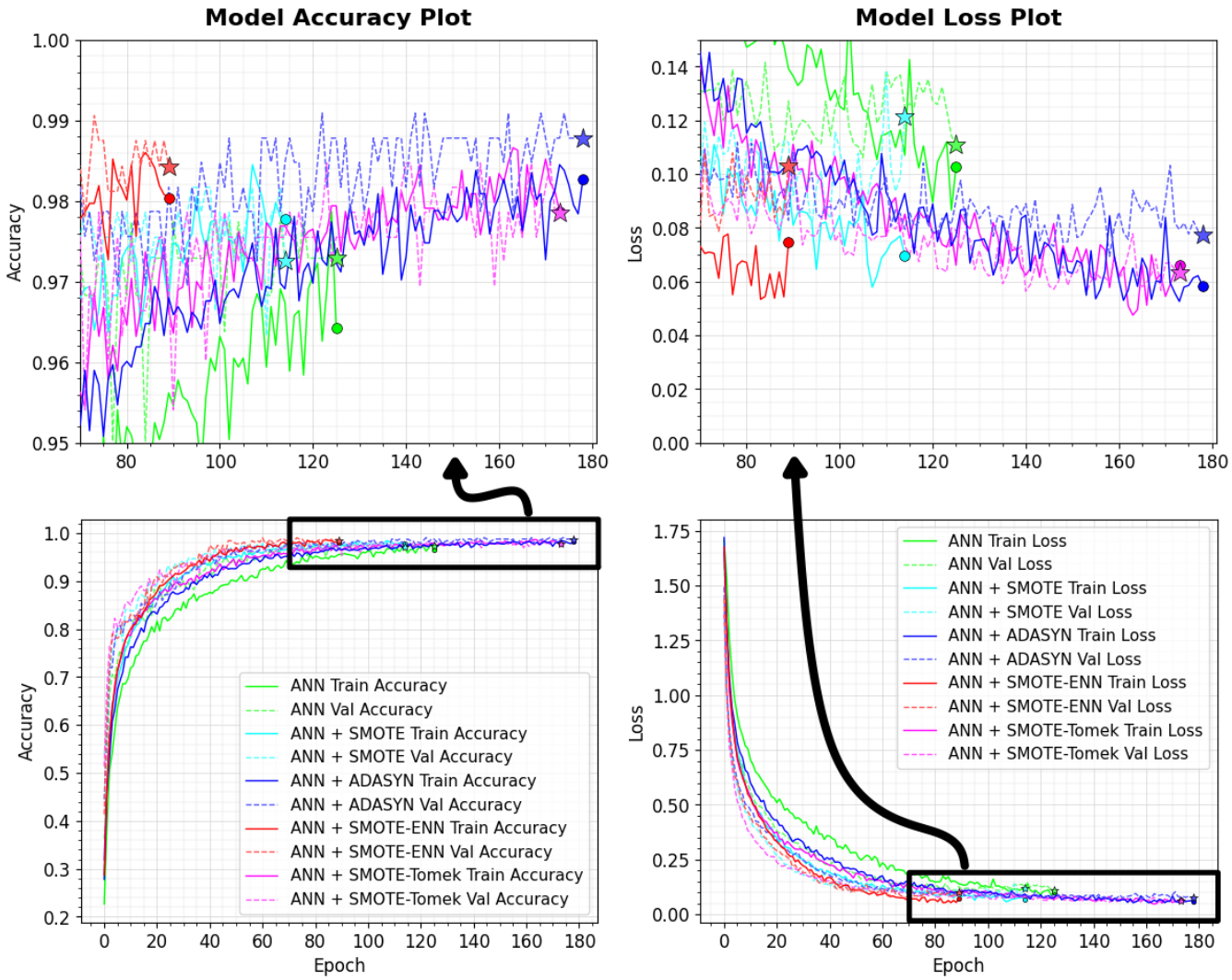


Fig. 7. Artificial Neural Network Model Accuracy and Loss Plot.

production environment.

VI. Results

Based on the model training results, the plots of accuracy and loss for the ANN model with various re-sampling methods, displayed per epoch in Fig. 7, are presented. The performance evaluation, measured using accuracy and loss function metrics, is also detailed in Table III. Several machine learning models, including Logistic Regression, Naïve Bayes, Support Vector Machine, and Adaptive Boosting, achieved relatively low accuracy and validation accuracy, even after the application of re-sampling techniques. On the other hand, Decision Tree and Random Forest obtained relatively high accuracy and validation accuracy, surpassing other Machine Learning algorithms. Moreover, the Decision Tree algorithm with the SMOTE-ENN method yields perfect results, achieving 100% accuracy in both training and validation, as well as a loss function of 0 in both training and validation.

For the Deep Learning model using an ANN architecture, the last epoch of 126 achieved an accuracy of 96.42% in training and 97.28% in validation, as well as a loss function of 0.1027 in training and 0.1109 in validation. Meanwhile, ANN + ADASYN stopped at epoch 179 and achieved the highest accuracy of 98.27% in training and 98.78% in validation, along with a loss function of 0.0583 in training and 0.0775 in validation. Meanwhile, the results of training the ANN + SMOTE model, which stopped at epoch 115, yielded an accuracy of 97.77% in training and 97.25% in validation, along with a loss function of 0.0696 in training

TABLE III
Comparison of Model Training and Validation Accuracy and Loss

Model	Re-sampling	Last Epoch	Accuracy	Validation Accuracy	Loss	Validation Loss	Learning Rate
Logistic Regression (LR)	-	-	73.1590%	69.2308%	0.779712	0.807603	-
	ADASYN	-	71.4953%	71.9512%	0.751692	0.768067	-
	SMOTE	-	73.7316%	74.6177%	0.716195	0.678704	-
	SMOTE-ENN	-	73.8456%	73.9812%	0.705439	0.710769	-
	SMOTE-Tomek	-	76.1441%	74.0061%	0.710507	0.750583	-
Decision Tree (DT)	**-**	**-**	**100%**	**99.0950%**	**0**	**0.326187**	**-**
	ADASYN	**-**	**100%**	**98.7805%**	**0**	**0.439557**	**-**
	SMOTE	**-**	**100%**	**99.0826%**	**0**	**0.330676**	**-**
	SMOTE-ENN	**-**	**100%**	**100%**	**0**	**0**	**-**
	SMOTE-Tomek	**-**	**100%**	**98.4709%**	**0**	**0.551126**	**-**
Naïve Bayes (NB)	-	-	57.6307%	49.3213%	1.986547	2.200764	-
	ADASYN	-	51.2221%	52.7439%	2.312288	2.375088	-
	SMOTE	-	51.6013%	56.5749%	2.029027	1.742625	-
	SMOTE-ENN	-	55.0055%	52.9781%	1.989398	1.986201	-
	SMOTE-Tomek	-	53.9099%	53.5168%	2.056124	2.305578	-
Random Forest (RF)	**-**	**-**	**100%**	**99.0950%**	**0.016213**	**0.049468**	**-**
	ADASYN	**-**	**100%**	**99.6951%**	**0.014299**	**0.039042**	**-**
	SMOTE	**-**	**100%**	**99.6942%**	**0.012791**	**0.037692**	**-**
	SMOTE-ENN	**-**	**100%**	**99.6865%**	**0.007926**	**0.025254**	**-**
	SMOTE-Tomek	**-**	**100%**	**99.0826%**	**0.011800**	**0.042424**	**-**
Support Vector Machine (SVM)	-	-	87.3533%	87.7828%	0.370035	0.386369	-
	ADASYN	-	89.1804%	87.8049%	0.304728	0.346465	-
	SMOTE	-	89.6006%	86.2385%	0.272264	0.347936	-
	SMOTE-ENN	-	90.5430%	88.4013%	0.229951	0.296290	-
	SMOTE-Tomek	-	89.9099%	88.0734%	0.268388	0.339691	-
K-Nearest Neighbors (KNN)	-	-	91.7823%	90.4977%	0.145812	1.576748	-
	ADASYN	-	94.9317%	93.9024%	0.091279	0.751690	-
	SMOTE	-	95.0702%	90.5199%	0.089210	1.099186	-
	SMOTE-ENN	-	95.7518%	94.0439%	0.076769	0.749633	-
	SMOTE-Tomek	-	95.0991%	92.9664%	0.085947	1.072943	-
Adaptive Boosting (AdaBoost)	-	-	38.3671%	38.4615%	1.752383	1.752381	-
	ADASYN	-	33.3214%	33.2317%	1.752558	1.752560	-
	SMOTE	-	50.0180%	49.8471%	1.753136	1.753141	-
	SMOTE-ENN	-	50.4987%	50.4702%	1.753114	1.753116	-
	SMOTE-Tomek	-	33.3333%	33.0275%	1.752546	1.752546	-
Artificial Neural Network (ANN)	-	126	96.4248%	97.2851%	0.102722	0.110863	0.001
	ADASYN	**179**	**98.2746%**	**98.7805%**	**0.058273**	**0.077462**	**0.001**
	SMOTE	115	97.7690%	97.2477%	0.069598	0.121554	0.001
	SMOTE-ENN	**90**	**98.0421%**	**98.4326%**	**0.074591**	**0.103261**	**0.001**
	SMOTE-Tomek	174	97.8708%	97.8593%	0.066034	0.063263	0.001

and 0.1216 in validation.

Among the models, the results of model training with a combination of oversampling and undersampling techniques with ANN + SMOTE-ENN stopped at the fewest epochs among other models, 90 epochs, obtained an accuracy of 98.04% in training and 98.43% in validation, slightly lower than ANN + ADASYN, with a loss function of 0.0746 in training and 0.1033 in validation. Meanwhile, ANN + SMOTE-Tomek stopped at 174 epochs, obtained an accuracy of 97.87% in training and 97.86% in validation, and a loss function of 0.0660 in training and 0.0633 in validation.

From Table IV, it is found that the ANN + ADASYN method is the best-performing model, with a significant increase in accuracy when compared to only using ANN without the over/undersampling method, by 1.85% in training accuracy and 1.50% in validation accuracy, with a difference in loss of 0.0444 and 0.0334 in validation loss. ANN + SMOTE-ENN is the second-best-performing model, with an accuracy increase of 1.62% and a validation accuracy of 1.15%. It achieves a lower loss of 0.0281 and 0.0076 for the validation loss. The other two methods, SMOTE and SMOTE-Tomek, did not have a better effect than the previous two methods.

The Decision Tree and Random Forest algorithms did not show significant improvements because, by using only the original unbalanced dataset, they achieved perfect training accuracy. Furthermore, only slight improvements in validation accuracy were observed, with 0.91% using DT + SMOTE-ENN, 0.6% using RF + ADASYN and RF + SMOTE, and 0.59% using RF + SMOTE-ENN. The same improvement also occurs in the loss and validation loss, which both show improvements when using the proposed re-sampling techniques.

TABLE IV
DIFFERENCE OF THE MODEL TRAINING AND VALIDATION PERFORMANCE RESULTS WITH AND WITHOUT THE PROPOSED METHODS

Method	Accuracy	Val. Accuracy	Loss	Val. Loss
DT + ADASYN	+0%	-0,32%	-0.0	+0,11337
DT + SMOTE	+0%	-0,01%	-0.0	+0,004489
DT + SMOTE-ENN	**+0%**	**+0,91%**	**-0.0**	**-0,326187**
DT + SMOTE-Tomek	+0%	-0,62%	-0.0	+0,224939
RF + ADASYN	**+0%**	**+0,60%**	**-0,001914**	**-0,010426**
RF + SMOTE	**+0%**	**+0,60%**	**-0,003422**	**-0,011776**
RF + SMOTE-ENN	**+0%**	**+0,59%**	**-0,008287**	**-0,024214**
RF + SMOTE-Tomek	+0%	-0,01%	-0,004413	-0,007044
ANN + ADASYN	**+1.85%**	**+1.50%**	**-0.044449**	**-0.033401**
ANN + SMOTE	+1.34%	-0.04%	-0.033124	+0.010691
ANN + SMOTE-ENN	**+1.62%**	**+1.15%**	**-0.028131**	**-0.007602**
ANN + SMOTE-Tomek	+1.45%	+0.57%	-0.036688	-0.0476

After training, the models are evaluated on both the original test dataset and a modified test dataset created using oversampling and undersampling techniques. Table V presents the detailed results of the comparative evaluation of model testing.

With the Machine Learning model, several algorithms such as Decision Tree, Naïve Bayes, Random Forest, Support Vector Machine, and Adaptive Boosting achieved their best performance when paired with the SMOTE-ENN technique. For instance, the Decision Tree with SMOTE-ENN achieved an accuracy of 98.75%, a loss of 0.4505, and precision, recall, and F1-score of 99%. Likewise, Random Forest with SMOTE-ENN yielded the highest overall performance, with a testing accuracy of 99.37%, loss of 0.0309, and equally strong precision, recall, and F1-score of 99%.

TABLE V
COMPARISON OF MODEL TESTING EVALUATIONS

Method	Accuracy	Loss	Precis-ion	Recall	F1-score
LR	78.38%	0.7286	79%	78%	78%
LR + ADASYN	64.02%	0.8224	62%	64%	62%
LR + SMOTE	78.66%	0.7173	80%	79%	78%
LR + SMOTE-ENN	76.87%	0.7040	78%	77%	76%
LR + SMOTE-Tomek	76.22%	0.7213	77%	76%	76%
DT	98.20%	0.6494	98%	98%	98%
DT + ADASYN	97.56%	0.8791	98%	98%	98%
DT + SMOTE	98.17%	0.6593	98%	98%	98%
DT + SMOTE-ENN	**98.75%**	**0.4505**	**99%**	**99%**	**99%**
DT + SMOTE-Tomek	96.95%	1.0989	97%	97%	97%
NB	52.25%	2.0400	59%	52%	47%
NB + ADASYN	50.00%	2.3661	48%	50%	43%
NB + SMOTE	53.05%	1.7813	65%	53%	49%
NB + SMOTE-ENN	55.00%	1.7367	73%	55%	51%
NB + SMOTE-Tomek	53.05%	2.3603	63%	53%	47%
RF	98.20%	0.0630	98%	98%	98%
RF + ADASYN	98.17%	0.0688	98%	98%	98%
RF + SMOTE	97.56%	0.0716	98%	98%	98%
RF + SMOTE-ENN	**99.37%**	**0.0309**	**99%**	**99%**	**99%**
RF + SMOTE-Tomek	96.34%	0.0845	96%	96%	96%
SVM	87.39%	0.3717	88%	87%	87%
SVM + ADASYN	84.15%	0.4570	84%	84%	83%
SVM + SMOTE	87.19%	0.3617	88%	87%	87%
SVM + SMOTE-ENN	88.75%	0.2490	90%	89%	88%
SVM + SMOTE-Tomek	87.81%	0.3683	88%	88%	88%
KNN	87.39%	1.7675	88%	87%	87%
KNN + ADASYN	92.68%	0.5696	93%	93%	93%
KNN + SMOTE	92.68%	1.1894	93%	93%	93%
KNN + SMOTE-ENN	92.50%	0.5360	93%	93%	92%
KNN + SMOTE-Tomek	93.29%	1.6153	94%	93%	93%
AdaBoost	37.84%	1.7524	16%	38%	22%
AdaBoost + ADASYN	33.54%	1.7525	11%	34%	17%
AdaBoost + SMOTE	50.00%	1.7531	30%	50%	36%
AdaBoost + SMOTE-ENN	50.63%	1.7531	31%	51%	36%
AdaBoost + SMOTE-Tomek	33.54%	1.7525	14%	34%	19%
ANN	93.69%	0.0631	94%	94%	94%
ANN + ADASYN	96.34%	0.0366	97%	96%	96%
ANN + SMOTE	95.73%	0.0427	96%	96%	96%
ANN + SMOTE-ENN	**99.38%**	**0.0062**	**99%**	**99%**	**99%**
ANN + SMOTE-Tomek	95.12%	0.0488	96%	95%	95%

Other models also demonstrated improved results using various resampling methods. Logistic Regression performed best with SMOTE, and K-Nearest Neighbor excelled with SMOTE-Tomek. However, it is essential to note that specific resampling techniques had a negative impact on

performance in some cases. For example, Logistic Regression with ADASYN, Decision Tree with SMOTE-Tomek, Naïve Bayes with ADASYN, Random Forest with SMOTE-Tomek, SVM with ADASYN, and AdaBoost with ADASYN experienced a drop in testing accuracy, precision, recall, f1-score, and an increase in loss compared to their performance on the original dataset.

In contrast to traditional machine learning models, the Deep Learning model, which utilizes an Artificial Neural Network (ANN), consistently demonstrates strong performance across all resampling strategies. As shown in Table V, each technique — ADASYN, SMOTE, SMOTE-ENN, and SMOTE-Tomek — contributes to performance improvement in the ANN model. This indicates that the neural network is highly adaptable to resampled data and can learn from modified class distributions more effectively than some classical algorithms.

Among these, the ANN combined with SMOTE-ENN emerges as the best-performing deep learning model, achieving a testing accuracy of 99.38%, a loss function of just 0.0062, and precision, recall, and f1-score of 99% each. These results suggest not only a very high correct classification rate but also minimal prediction error, equivalent to only one misclassified instance in the entire test set. The consistent 99% across all evaluation metrics reflects a balanced and highly reliable model. These findings reinforce the potential of deep learning, particularly when combined with advanced resampling methods, in detecting complex patterns commonly found in medical datasets.

Beyond just accuracy, metrics such as precision, recall, and F1-score are critical in the context of medical diagnosis. In diseases like celiac disease, which are often underdiagnosed or misdiagnosed, a high recall is particularly vital. This means the model correctly identifies most patients who actually have the disease, minimizing false negatives. Failing to detect a true positive could lead to prolonged patient suffering or complications due to untreated gluten intolerance. At the same time, high precision ensures that those diagnosed by the model are genuinely at risk, reduces unnecessary anxiety, costly follow-up tests, and possible dietary restrictions imposed on those without the condition. The f1-score provides a harmonic mean between these two metrics, offering a single, balanced view of a model's diagnostic capability.

A clear distinction is made from Table VI, which presents the change in testing performance for each model when enhanced with various oversampling and undersampling techniques. Among the machine learning models, Decision Tree and Random Forest, paired with SMOTE-ENN, achieved slight improvements in test accuracy, of +0.55% and +1.17%, respectively. Despite near-perfect training and validation, both models showed only modest improvement and limited generalization on test data. This suggests overfitting and highlights that high training accuracy does not always translate to strong testing performance, particularly in sensitive, real-world medical diagnoses.

By comparison, the deep learning model (ANN) exhibited more substantial improvements across all evaluation metrics when combined with SMOTE-ENN. The ANN + SMOTE-ENN model recorded the highest increase in testing accuracy (5.69%) and the largest decrease in loss (0.0569), indicating a significant reduction in false predictions. Additionally, it achieved a 5% improvement in precision, recall, and F1-score. This consistency across all metrics suggests that this model not only classifies correctly more often but also makes far fewer critical errors, such as missing true celiac cases (false negatives) or incorrectly identifying healthy individuals as positive (false positives).

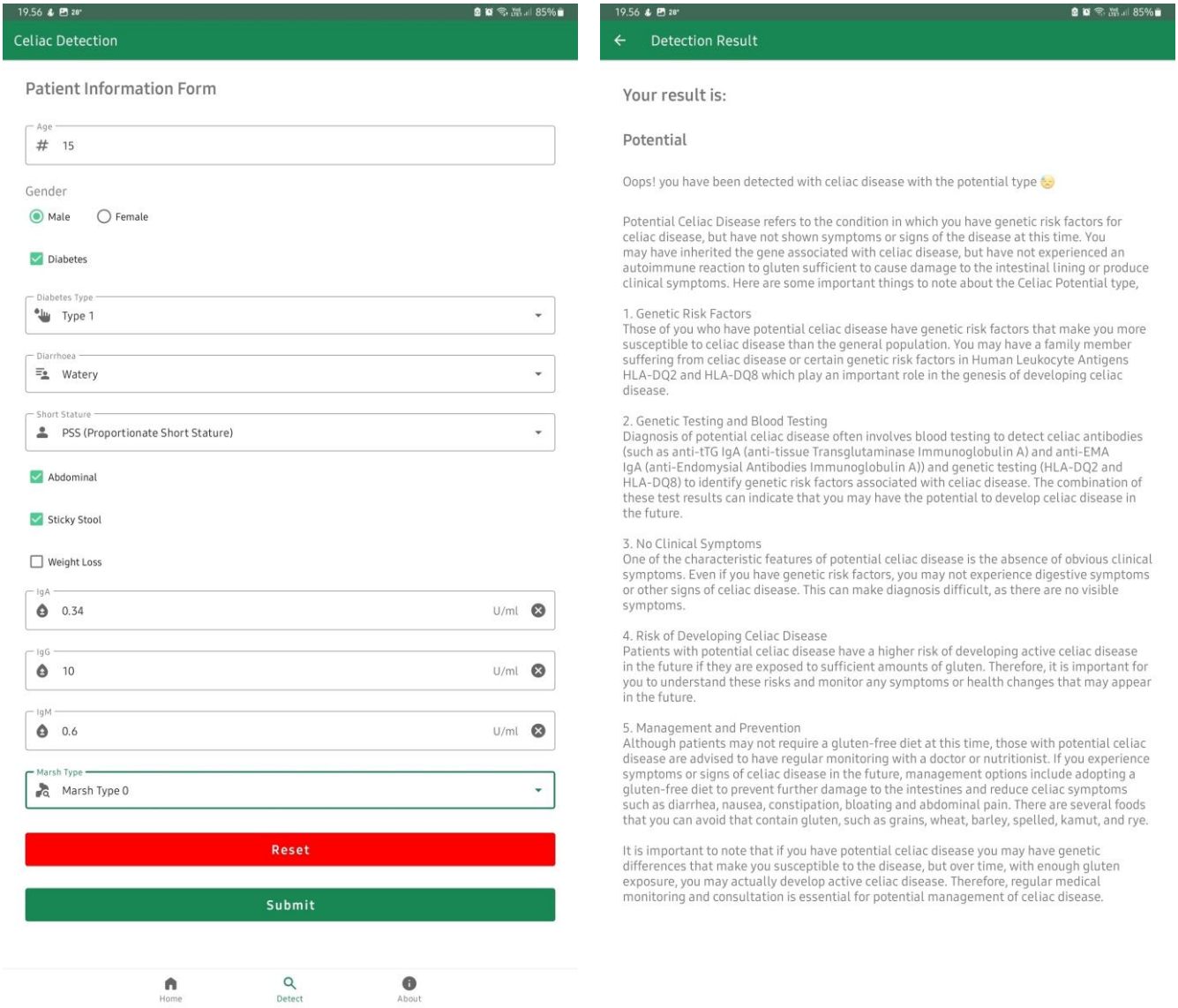
While the Random Forest + SMOTE-ENN model came close in terms of overall performance, achieving 99.37% accuracy and a loss of 0.0309, compared to 99.38% accuracy and a loss of 0.0062 for ANN + SMOTE-ENN, it is important to note that the ANN still performed marginally better in both classification correctness and error reduction. Both models achieved identical precision, recall, and F1-score (99%), which indicates excellent class-wise balance. However, the ANN + SMOTE-ENN model achieved the lowest testing loss among all models evaluated, suggesting that its predictions were not only accurate but also made with higher confidence and better probability calibration.

These results affirm that combining oversampling (SMOTE) with undersampling (ENN) can enhance model generalization better than oversampling alone, particularly when applied to a deep learning model like ANN. While the machine learning models showed only modest improvements, typically around 1% in accuracy and other evaluation metrics, the deep learning model demonstrated the highest gains across all metrics. This highlights the strong potential of deep learning, when supported by appropriate resampling techniques, to deliver robust and generalizable solutions in medical classification tasks such as celiac disease detection.

TABLE VI
DIFFERENCE OF THE MODEL TESTING EVALUATION RESULTS WITH AND WITHOUT THE PROPOSED METHODS

Method	Accuracy	Loss	Precis-ion	Recall	F1-score
DT + ADASYN	-0.64%	+0.2297	+0%	+0%	+0%
DT + SMOTE	-0.03%	+0.0099	+0%	+0%	+0%
DT + SMOTE-ENN	**+0.55%**	**-0.1989**	**+1%**	**+1%**	**+1%**
DT + SMOTE-Tomek	-1.25%	+0.4495	-1%	-1%	-1%
RF + ADASYN	-0.03%	+0.0058	+0%	+0%	+0%
RF + SMOTE	-0.64%	+0.0086	+0%	+0%	+0%
RF + SMOTE-ENN	**+1.17%**	**-0.0321**	**+1%**	**+1%**	**+1%**
RF + SMOTE-Tomek	-1.86%	+0.0215	-2%	-2%	-2%
ANN + ADASYN	+2.65%	-0,0265	+3%	+2%	+2%
ANN + SMOTE	+2.04%	-0,0204	+2%	+2%	+2%
ANN + SMOTE-ENN	**+5.69%**	**-0,0569**	**+5%**	**+5%**	**+5%**
ANN + SMOTE-Tomek	+1.43%	-0,0143	+2%	+1%	+1%

To enable deployment on mobile devices, the trained model was optimized by converting it from its original HDF5 (Hierarchical Data Format) file size of 5.78 MB to a lightweight TensorFlow Lite (TF-Lite) format, resulting in a reduced size of 1.91 MB. This conversion significantly improves model efficiency and inference speed, making it suitable for real-time predictions in resource-constrained environments, such as smartphones. The optimized ANN + SMOTE-ENN model was successfully integrated into the Android-based celiac disease detection application. The system enables users to input new patient data and receive



(a) Detect Page (b) Result Page

Fig. 8. Implementation of Artificial Neural Network + SMOTE-ENN Model in an Android-based Mobile Application.

immediate predictions regarding the presence or absence of celiac disease.

As illustrated in Fig. 8(a), the Android application features a user-friendly interface for inputting medical information relevant to the diagnosis. Fig. 8(b) displays the prediction result, which not only classifies the condition (e.g., positive or negative for celiac disease) but also provides a brief explanation of the disease type and personalized recommendations. This integration demonstrates the practical applicability of the proposed model in real-world healthcare settings, supporting early screening and personalized assistance for individuals who may be at risk of developing celiac disease.

VII. Conclusion and Future Work

This research investigated the impact of resampling techniques on improving celiac disease detection, focusing on methods such as ADASYN, SMOTE, SMOTE-Tomek, and SMOTE-ENN. Given the significant class imbalance in the dataset, these techniques were applied to enhance classification performance. Among various models, the Artificial Neural Network (ANN) combined with SMOTE-ENN delivered the best results, achieving a testing accuracy of 99.38% with a minimal loss of 0.0062. It also reached precision, recall, and F1-score of 99%, showing excellent class-wise prediction balance. This is especially crucial in medical diagnosis, where high recall helps minimize false negatives, ensuring true cases are not overlooked, and high precision reduces false positives, preventing unnecessary stress or treatment. The F1-score further confirms the model's balanced performance across both classes.

The superior results of SMOTE-ENN can be attributed to its two-step process, oversampling minority instances and then removing noisy majority class samples using Edited Nearest Neighbors. This not only balances the data but also enhances model robustness by reducing class overlap and noise. Compared to other resampling methods, SMOTE-ENN consistently led to better generalization on the test set. The final model was optimized into TensorFlow Lite format and successfully integrated into an Android application for real-time, user-friendly celiac disease screening,

demonstrating its practical applicability in mobile health technology.

For future work, expanding the dataset with more varied and clinically validated samples is essential to improve model generalization. Real-time monitoring of model performance, integration of user feedback, and mechanisms to address data drift over time should also be considered. These enhancements will support long-term reliability, ethical deployment, and broader adoption of AI-driven tools for early detection and personalized care in celiac disease and beyond.

REFERENCES

[1] B. Lebwohl and A. Rubio-Tapia, "Epidemiology, presentation, and diagnosis of celiac disease," *Gastroenterology*, vol. 160, no. 1, pp. 63–75, Jan. 2021, doi: 10.1053/j.gastro.2020.06.098.

[2] R. Tomer, S. Patiyal, A. Dhall, and G. P. S. Raghava, "Prediction of celiac disease associated epitopes and motifs in a protein.," *Front. Immunol.*, vol. 14, p. 1056101, 2023, doi: 10.3389/fimmu.2023.1056101.

[3] J. Vicnesh *et al.*, "Automated diagnosis of celiac disease by video capsule endoscopy using daisy descriptors," *J. Med. Syst.*, vol. 43, no. 6, p. 157, 2019, doi: 10.1007/s10916-019-1285-6.

[4] S. Syed *et al.*, "Assessment of machine learning detection of environmental enteropathy and celiac disease in children," *JAMA Netw. Open*, vol. 2, no. 6, pp. e195822–e195822, Jun. 2019, doi: 10.1001/jamanetworkopen.2019.5822.

[5] G. Perrotta and E. Guerrieri, "Celiac disease: definition, classification, historical and epistemological profiles, anatomopathological aspects, clinical signs, differential diagnosis, treatments and prognosis. proposed diagnostic scheme for celiac disease (dscnc)," *Arch. Clin. Gastroenterol.*, vol. 8, no. 1, pp. 008–019, Apr. 2022, doi: 10.17352/2455-2283.000106.

[6] L. Macedo, M. Catarino, C. Festas, and P. Alves, "Vulnerability in children with celiac disease: findings from a scoping review," *Children*, vol. 11, no. 6, p. 729, Jun. 2024, doi: 10.3390/children11060729.

[7] G. K. Makharia, "Celiac disease screening in southern and east asia," *Dig. Dis.*, vol. 33, no. 2, pp. 167–174, 2015, doi: 10.1159/000369537.

[8] H. Oktadiana, M. Abdullah, K. Renaldi, and N. Dyah, "Diagnosis dan tata laksana penyakit celiac," *J. Penyakit Dalam Indones.*, vol. 4, no. 3, p. 157, Sep. 2017, doi: 10.7454/jpdi.v4i3.131.

[9] S. Mohta, M. S. Rajput, V. Ahuja, and G. K. Makharia, "Emergence of celiac disease and gluten-related disorders in asia," *J. Neurogastroenterol. Motil.*, vol. 27, no. 3, pp. 337–346, Jul. 2021, doi: 10.5056/jnm20140.

[10] M. K. Sana, Z. M. Hussain, P. A. Shah, and M. H. Maqsood, "Artificial intelligence in celiac disease," *Comput. Biol. Med.*, vol. 125, p. 103996, Oct. 2020, doi: 10.1016/j.compbiomed.2020.103996.

[11] C.-A. Stoleru, E. H. Dulf, and L. Ciobanu, "Automated detection of celiac disease using machine learning algorithms," *Sci. Rep.*, vol. 12, no. 1, p. 4071, 2022, doi: 10.1038/s41598-022-07199-z.

[12] J. Win, J. Ho, T. Mangala, and J. Koul, "Celiac disease (coeliac disease)," Kaggle.com. Accessed: Jul. 22, 2024. [Online]. Available: https://kaggle.com/datasets/jackwin07/celiac-disease-coeliac-disease

[13] A. Amalia, M. S. Lydia, S. M. Hardi, and A. B. Jamesie, "Implementation of celiac disease detection using a website-based artificial neural network approach," in *2023 7th International Conference on Electrical, Telecommunication and Computer Engineering (ELTICOM)*, Medan: IEEE, Dec. 2023, pp. 134–138. doi: 10.1109/ELTICOM61905.2023.10443109.

[14] K. Ghosh *et al.*, "The class imbalance problem in deep learning," *Mach. Learn.*, vol. 113, no. 7, pp. 4845–4901, Jul. 2024, doi: 10.1007/s10994-022-06268-8.

[15] T. M. Alam *et al.*, "An efficient deep learning-based skin cancer classifier for an imbalanced dataset," *Diagnostics*, vol. 12, no. 9, p. 2115, Aug. 2022, doi: 10.3390/diagnostics12092115.

[16] J. Denholm *et al.*, "Multiple-instance-learning-based detection of coeliac disease in histological whole-slide images," *J. Pathol. Inform.*, vol. 13, p. 100151, 2022, doi: 10.1016/j.jpi.2022.100151.

[17] J. E. W. Koh *et al.*, "Automated interpretation of biopsy images for the detection of celiac disease using a machine learning approach," *Comput. Methods Programs Biomed.*, vol. 203, p. 106010, May 2021, doi: 10.1016/j.cmpb.2021.106010.

[18] J. Carreras, "Celiac disease deep learning image classification using convolutional neural networks," *J. Imaging*, vol. 10, no. 8, p. 200, Aug. 2024, doi: 10.3390/jimaging10080200.

[19] D.-H. Jeong, S.-E. Kim, W.-H. Choi, and S.-H. Ahn, "A comparative study on the influence of undersampling and oversampling techniques for the classification of physical activities using an imbalanced accelerometer dataset," *Healthcare*, vol. 10, no. 7, 2022, doi: 10.3390/healthcare10071255.

[20] L. K. Xin and N. binti A. Rashid, "Prediction of depression among women using random oversampling and random forest," in *2021 International Conference of Women in Data Science at Taif University (WiDSTaif)*, IEEE, Mar. 2021, pp. 1–5. doi: 10.1109/WiDSTaif52235.2021.9430215.

[21] R. Mohammed, J. Rawashdeh, and M. Abdullah, "Machine learning with oversampling and undersampling techniques: overview study and experimental results," in *2020 11th International Conference on Information and Communication Systems (ICICS)*, IEEE, Apr. 2020, pp. 243–248. doi: 10.1109/ICICS49469.2020.239556.

[22] M. M. Nishat *et al.*, "A comprehensive investigation of the performances of different machine learning classifiers with smote-enn oversampling technique and hyperparameter optimization for imbalanced heart failure dataset," *Sci. Program.*, vol. 2022, no. 1, pp. 1–17, Mar. 2022, doi: 10.1155/2022/3649406.

[23] M. F. Ijaz, M. Attique, and Y. Son, "Data-driven cervical cancer prediction model with outlier detection and over-sampling methods," *Sensors*, vol. 20, no. 10, 2020, doi: 10.3390/s20102809.

[24] M. Kumari and N. Subbarao, "A hybrid resampling algorithms smote and enn based deep learning models for identification of marburg virus inhibitors," *Future Med. Chem.*, vol. 14, no. 10, pp. 701–715, May 2022, doi: 10.4155/fmc-2021-0290.

[25] J. A. Adisa, S. Ojo, P. A. Owolawi, A. Pretorius, and S. O. Ojo, "The Effect of Imbalanced Data and Parameter Selection via Genetic Algorithm Long Short-Term Memory (LSTM) for Financial Distress Prediction," *IAENG International Journal of Applied Mathematics*, vol. 53, no. 3, pp. 796–809, 2023.

[26] K. Zhou, C. Zhang, Y. Yu, S. Cong, and X. Yue, "Improving Smote Technology for Credit Card Fraud Detection Category Imbalance Issues," *Engineering Letters*, vol. 31, no. 4, pp. 1780–1785, 2023.

[27] M. M. Nwe and K. T. Lynn, "Effective Resampling Approach for Skewed Distribution on Imbalanced Data Set," *IAENG International Journal of Computer Science*, vol. 47, no. 2, pp. 234–249, 2020.

[28] D. Kaul, H. Raju, and B. K. Tripathy, "Deep learning in healthcare," in *Deep Learning in Data Analytics: Recent Techniques, Practices and Applications*, D. P. Acharjya, A. Mitra, and N. Zaman, Eds., Cham: Springer International Publishing, 2022, pp. 97–115. doi: 10.1007/978-3-030-75855-4_6.

[29] M. J. Horry *et al.*, "COVID-19 detection through transfer learning using multimodal imaging data," *IEEE Access*, vol. 8, pp. 149808–149824, 2020, doi: 10.1109/ACCESS.2020.3016780.

[30] Y. Liu *et al.*, "Detecting diseases by human-physiological-parameter-based deep learning," *IEEE Access*, vol. 7, pp. 22002–22010, 2019, doi: 10.1109/ACCESS.2019.2893877.

[31] H. Abdel-Jaber, D. Devassy, A. Al Salam, L. Hidaytallah, and M. EL-Amir, "A review of deep learning algorithms and their applications in healthcare," *Algorithms*, vol. 15, no. 2, 2022, doi: 10.3390/a15020071.

[32] T. G. S. *et al.*, "An extension of synthetic minority oversampling technique based on kalman filter for imbalanced datasets," *Mach. Learn. with Appl.*, vol. 8, p. 100267, 2022, doi: 10.1016/j.mlwa.2022.100267.

[33] S. Goyal, "Handling class-imbalance with knn (neighbourhood) under-sampling for software defect prediction," *Artif. Intell. Rev.*, vol. 55, no. 3, pp. 2023–2064, 2022, doi: 10.1007/s10462-021-10044-w.

[34] H. Hartono and E. Ongko, "Avoiding overfitting dan overlapping in handling class imbalanced using hybrid approach with smoothed bootstrap resampling and feature selection," *JOIV Int. J. Informatics Vis.*, vol. 6, no. 2, p. 343, Jun. 2022, doi: 10.30630/joiv.6.2.985.

[35] G. Wei, W. Mu, Y. Song, and J. Dou, "An improved and random synthetic minority oversampling technique for imbalanced data," *Knowledge-Based Syst.*, vol. 248, p. 108839, Jul. 2022, doi: 10.1016/j.knosys.2022.108839.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 4, pp. 321–357, Jun. 2002, doi: 10.1613/jair.953.

[37] H. K. Fatlawi and A. Kiss, "An elastic self-adjusting technique for rare-class synthetic oversampling based on cluster distortion minimization in data stream," *Sensors*, vol. 23, no. 4, 2023, doi: 10.3390/s23042061.

[38] Y. Mi, "Imbalanced classification based on active learning smote,"

Res. J. Appl. Sci. Eng. Technol., vol. 5, no. 3, pp. 944–949, Jan. 2013, doi: 10.19026/rjaset.5.5044.

[39] S. Wang, Y. Dai, J. Shen, and J. Xuan, "Research on expansion and classification of imbalanced data based on smote algorithm," *Sci. Rep.*, vol. 11, no. 1, p. 24039, Dec. 2021, doi: 10.1038/s41598-021-03430-5.

[40] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *SIGKDD Explor. Newsl.*, vol. 6, no. 1, pp. 20–29, Jun. 2004, doi: 10.1145/1007730.1007735.

[41] B. Deepa and K. Ramesh, "Epileptic seizure detection using deep learning through min max scaler normalization," *Int. J. Health Sci. (Qassim).*, pp. 10981–10996, May 2022, doi: 10.53730/ijhs.v6nS1.7801 .